DOCTOR OF PHILOSOPHY

Genetic Epidemiology Studies of aspects of Diabetic Complications

Deshmukh, Harshal

Award date:
2014

Genetic Epidemiology Studies of aspects of Diabetic Complications

Dr Harshal Deshmukh

For the degree of PhD (Medicine)

University of Dundee

September 1st 2014

Table of Contents

Declaration of the Candidate x

Acknowledgements xi

Publications during PhD xii

Thesis Summary 14

I. Introduction 16

II. Definition of Diabetic Nephropathy 16

III. Concept of Chronic Kidney Disease (CKD) in Diabetes 18

IV. Concept of Diabetic Kidney Disease 19

V. Role of Retinopathy in Diagnosis of DKD 19

VI. Natural History of Diabetic Nephropathy and Nephropathy Phenotypes 20

VII. Current Understanding of Key Aspects of Histology and Pathogenesis of DKD including some Key Pathways 22

1. Histopathology of Diabetic Kidney Disease 22
2. Physiological Changes in DKD 23
3. Role of Cytokines and Vascular Growth Factors in DKD 25
4. Role of AGE-RAGE Axis in DKD 26
5. Role of Oxidative Stress DKD 27
6. Role of Lipids in Pathogenesis of DKD 27

VIII. Heritability of diabetic nephropathy 28

7. Familial Clustering of Diabetic Nephropathy 28
8. Heritability Estimates for Diabetic Nephropathy 30

IX. Animal Models for Diabetic Kidney Disease 31

X. Candidate Gene Studies, Sib pair Studies and GWAS Studies for Kidney Disease 32

9. Literature Review Candidate Gene Studies for Diabetic Nephropathy and CKD in Diabetes 32
10. Angiotensin Converting Enzyme gene (ACE) 33
11. Nitric Oxide Synthase 3 (NOS3) Gene 34
12. Apolipoprotein E (APOe) 36
13. TGFβ1 (Transforming growth factor beta 1) 36
14. Candidate Gene Studies Summary 37

15. Literature Review of Genome-wide Sib-pair studies for DKD 38
16. Genome-wide Association Studies for Diabetic Nephropathy 41
17. GWAS for T1D Nephropathy in Genetics of Kidneys in Diabetes Collection (GoKinD) 41
18. GWAS for Type 1 Diabetic Nephropathy in American European Population (Genie Study) 42
19. GWAS for Type 2 Diabetic Nephropathy in African American Population (McDnough 2010) 43
20. GWAS for DKD stratified by MYH9 and APOL1 Risk Variants 45
21. Summary of Genetic Epidemiology studies for Diabetic Nephropathy 48
Aims 51
Methods 54
I. Datasets used in the thesis 54
II. DARTS and Go-DARTS 54
22. CHI-Master Index and SMRO1 and GRO data 54
23. Laboratory Data for eGFR and Albuminuria (ACR and MA) and Renal Phenotypes 56
24. Genotype Data 61
III. CARDS Clinical Trial 63
25. CARDS Phenotype Data 64
26. CARDS Genotype Data 64
IV. EURODIAB Study 66
27. EURODIAB Phenotype data 66
28. EURODIAB Genotype data 67
29. EURODIAB Genotype data quality control 67
V. Statistical Analysis 69
VI. Genotype Imputation 69
VII. Haplotype Inference 69
VIII. Imputation with IMPUTEv2 70
IX. Cluster Computing 70
X. Genetic Associations with SNPTEST 71
XI. Principal Component Analysis using EIGENSOFT 72
XII. Example of Principal Component Analysis using the EURODIAB Dataset 73
XIII. Meta-analysis 74
XIV. Fixed Effects Meta-analysis 75

XV. Random Effects Meta-analysis 75
XVI. Specific Methods for Heritability Estimation using GCTA 76
XVIII. Specific Methods for sRAGE GWAS 77
30. CARDS Dataset 77
31. Go-DARTS Dataset 78
32. Conditional Analysis of RAGE Gene in CARDS Dataset 79
33. Chip-based Heritability Estimation and Genome-wide Joint Conditional Analysis using Genome-Complex Trait Analysis (GCTA) 80
34. Power Calculation 81
35. Meta-analysis 82
XIX. Specific Methods for CKDGen Replication 82
36. Power calculation 84
37. Genetic Risk Score 84
38. Model selection 85
XX Specific methods for SUMMIT GWAS analysis for DKD 85
39. Phenotype Definitions 85
40. Imputation 87
41. Analysis details 87
42. SNPTEST Runs 88
43. Meta-analysis 88
44. Power Calculation for SUMMIT DKD studies 88
Results 90
I. Estimating Heritability of Various DN Phenotypes 90
I. Association of known SNPs for upstream risk factors for DN and their association with Diabetic Kidney Disease 95
III. GWAS for biomarker for DN (sRAGE) 100
IV. Replication of SNPs in CKDgen I consortium and contribution to CKDgen II consortium 115
45. Results 1: Replication of CKDgen Consortium I 115
46. Results II: Contribution to CKDgen consortium II 123
V. GWAS for DKD in EURODIAB and GO-DARTS dataset 127
VI. Meta-analysis of GWAS data with other datasets 136
Discussion 154
I. Estimating chip-based heritability of various DN phenotypes 154

II. GWAS for upstream risk factors of DKD (BP and HbA1c) OR doing Genetic risk scores with known loci and testing their association with DN phenotypes 157
III. VI. Replication of CKDgen Consortium eGFR loci Hits 161
IV. GWAS for biomarker for DN (sRAGE) 163
V. Meta-analysis of GWAS data with other DKD datasets 168
47. Novel Findings for DKD in Type 1 and Type 2 Diabetes 169
48. Replication of Previous Results 172
49. Limitations 172
50. Heterogeneity of DKD phenotype 172
51. Cross-sectional study design 173
52. Additional Confounding factors 173
53. Generalizability and Effect sizes 174
54. Functional analysis 174
55. Future Directions 175
References 179

List of Tables and Figures

Table 1 Definitions of the abnormalities of albumin excretion 17
Table 2: Likelihood of DKD according to Staging by eGFR and albuminuria level[6] 20
Table 3 Nephropathy phenotypes and corresponding hemodynamic and pathologic changes 22
Table 4: Top Candidate Genes for Diabetic Nephropathy 33
Table 5: Genome wide Association studies using Sib Pair Designs for Type 1 and Type 2 Diabetic Nephropathy 40
Table 6: Genome-wide Association Study for Diabetic Nephropathy 47
Table 7: Comparison of albuminuria status at baseline versus the albuminuria status at the last follow up using Macro-albumin (MA) test result 59
Table 8: Comparison of Albuminuria status at last follow-up using Albumin Creatinine ratio and Macro-albumin test 60
Table 9: Concordance rate between the GWAS sNPS common to the prior candidate SNP EURODIAB data 68
Table 10: Testing the GWAS data in well characterised variants and traits 68
Table 11: Cut-off used for Phenotype definitions in SUMMIT GWAS analysis of DKD 86
Table 12: Power calculations for T2D-DKD in SUMMIT studies 89
Table 13: Power calculations for T1D-DKD in SUMMIT studies 89
Table 14: Demographic Characteristics of the Go-DARTS Population 92
Table 15: Narrow –Sense Heritability of DN Phenotypes using GCTA 93
Table 16: Bivariate analysis of age and sex standardized residuals of average eGFR and average systolic blood pressure during the study period 94
Table 17: Bivariate analysis of age and sex standardized residuals of average eGFR and average HbA1c during the study period 94
Table 18 : Replication of SNPs associated with BP in Type 2 diabetes Go-DARTS (n=7276) (Total SNPS~ 60) 97
Table 19: Replication of SNPs associated with HbA1c in Type 2 diabetes Go-DARTS (n=7276) (Total SNPS~ 25) 98
Table 20: Genetic risk scores for BP (~60 SNPs), HbA1C (~25 SNPs) and its association with DKD 99
Table 21: Demographics and Baseline Characteristics of the Three Datasets 103
Table 22: Distribution of raw sRAGE values in CARDS and GO-DARTS datasets 104
Table 23: GWAS Meta-analysis for sRAGE in CARDS and GO-DARTS cohorts 106
Table 24 Genome-wide conditional analysis of CARDS-Go-DARTS dataset 107

Table 25 Replication of the top hits from CARDS- Go-DARTS meta-analysis in Type 1 and Type 2 Diabetes cohorts....107
Table 26: Reported eGFR associated loci in Kottgen A et al paper....118
Table 27: Demographic characteristics of the study population (Go-DARTS data genotyped on Affymetrix platform)....119
Table 28: Baseline characteristics of cases with progression to CKD Stage 3B CKD and Controls....119
Table 29: Association of the known SNPs with baseline eGFR, eGFR stratified by albuminuria status and time to stage 3B CKD....120
Table 30: Association of eGFR loci with Albuminuria in the Go-DARTS Cohort....121
Table 31: Replication for 35 SNPS for CKD 60 phenotype (N= 427 cases and 2466 controls) in Go-DARTS cohort....124
Table 32: Replication for 35 SNPS for CKD 45 phenotype (N= 195 cases and 2466 controls) in Go-DARTS cohort....125
Table 33: Replication for 35 SNPS for eGFR phenotype (N=2893) in Go-DARTS cohort 126
Table 34: Sample Size for T2D-DKD phenotypes....134
Table 35: Sample Size for T1D-DKD phenotypes....134
Table 36: Top GWAS hit for each of the DKD phenotypes in EURODIAB and Go-DARTS dataset....135
Table 37: Baseline characteristics of Type 1 Diabetes Cohort....138
Table 38: Baseline characteristics of Type 2 diabetes cohort....139
Table 39: Meta-Analysis of CKD phenotype in SUMMIT consortium in patients with Type 1 Diabetes....140
Table 40: Meta-Analysis of CKD-DN phenotype in SUMMIT consortium in patients with Type 1 Diabetes....141
Table 41: Meta-Analysis of ESRD phenotype in SUMMIT consortium in patients with Type 1 Diabetes....142
Table 42: Meta-Analysis of Macro-ESRD phenotype in SUMMIT consortium in patients with Type 1 Diabetes....143
Table 43: Meta-Analysis of Micro phenotype in SUMMIT consortium in patients with Type 1 Diabetes....144
Table 44: Meta-Analysis of DN phenotype in SUMMIT consortium in patients with Type 1 Diabetes....145
Table 45: Meta-Analysis of CKD phenotype in SUMMIT consortium in patients with Type 2 Diabetes....146
Table 46: Meta-Analysis of CKD-DN phenotype in SUMMIT consortium in patients with Type 2 Diabetes....147
Table 47: Meta-Analysis of ESRD phenotype in SUMMIT consortium in patients with Type 2 Diabetes....148
Table 48: Meta-Analysis of Macro-ESRD phenotype in SUMMIT consortium in patients with Type 2 Diabetes....149
Table 49: Meta-Analysis of Micro phenotype in SUMMIT consortium in patients with Type 2 Diabetes....150
Table 50: Meta-Analysis of DN phenotype in SUMMIT consortium in patients with Type 2 Diabetes....151

Table 51: Replication of reported loci associated with diabetic kidney disease in SUMMIT meta-analysis..... 152

Figure 1: DARTS and Go-DARTS study 56
Figure 2: Principal Component Analysis in EURODIAB Dataset 74
Figure 3: Distribution of Log10-sRAGE in CARDS 104
Figure 4: Distribution of Log10-sRAGE in Go-DARTS dataset 105
Figure 5: QQ plot of GWAS meta-analysis for sRAGE in CARDS and Go-DARTS dataset 108
Figure 6: Manhattan plot of GWAS meta-analysis for sRAGE in CARDS and Go-DARTS dataset 109
Figure 7: Regional Association plot RAGE-HLA gene region – association with circulating sRAGE levels 110
Figure 8: QQ plot for GWAS of sRAGE levels of CARDS Perlegen dataset 111
Figure 9: QQ plot for GWAS of sRAGE levels of Go-DARTS affymetrix dataset 112
Figure 10: QQ plot for GWAS of sRAGE levels of Go-DARTS Illumina dataset 113
Figure 11: Result of LD based pruning of the top 22 markers associated with sRAGE levels in CARDS and Go-DARTS meta-analysis 114
Figure 12 : Linkage Disequilibrium plot of the 16 known SNPs included in the study looking for association with eGFR 122
Figure 13: QQ-Plots of DN phenotype in EURODIAB dataset 129
Figure 14: QQ-Plots of CKD phenotype in EURODIAB dataset 129
Figure 15: QQ-Plots of CKD-DN phenotype in EURODIAB dataset 129
Figure 16: QQ-Plots of Macro-albuminuria phenotype in EURODIAB dataset 130
Figure 17: QQ-Plots of Micro-albuminuria phenotype in EURODIAB dataset 130
Figure 18: QQ-Plots of ESRD phenotype in EURODIAB dataset 131
Figure 19: QQ-Plots of DN phenotype in GO-DARTS dataset 131
Figure 20: QQ-Plots of CKD phenotype in GO-DARTS dataset 131
Figure 21: QQ-Plots of CKD-DN phenotype in GO-DARTS dataset 132
Figure 22: QQ-Plots of Macro-albuminuria phenotype in GO-DARTS dataset 132
Figure 23: QQ-Plots of Micro-albuminuria phenotype in GO-DARTS dataset 133
Figure 24: QQ-Plots of ESRD phenotype in GO-DARTS dataset 133
Figure 25: Manhattan Plot and QQ-plot SUMMIT CKD Phenotype T1D 140
Figure 26: Manhattan Plot and Q-Q plot SUMMIT CKD-DN Phenotype Type 1 Diabetes 141
Figure 27: Manhattan Plot and Q-Q plot SUMMIT ESRD Phenotype Type 1 Diabetes 142
Figure 28: Manhattan Plot and Q-Q plot SUMMIT Macro-ESRD Phenotype Type 1 Diabetes 143
Figure 29: Manhattan Plot and Q-Q plot SUMMIT Micro-albuminuria Phenotype Type 1 Diabetes 144
Figure 30: Manhattan Plot and Q-Q plot SUMMIT DN Phenotype Type 1 Diabetes 145
Figure 31: Manhattan Plot and Q-Q plot SUMMIT CKD Phenotype Type 2 Diabetes 146
Figure 32: Manhattan Plot and Q-Q plot SUMMIT CKD-DN Phenotype Type 2 Diabetes 147
Figure 33: Manhattan Plot and Q-Q plot SUMMIT ESRD Phenotype Type 2 Diabetes 148
Figure 34: Manhattan Plot and Q-Q plot SUMMIT Macro-ESRD Phenotype Type 2 Diabetes 149
Figure 35: Manhattan Plot and Q-Q plot SUMMIT Micro-albuminuria Phenotype Type 2 Diabetes 150
Figure 36: Manhattan Plot and Q-Q plot SUMMIT-DN Phenotype Type 2 Diabetes 151

Statement of Authorship

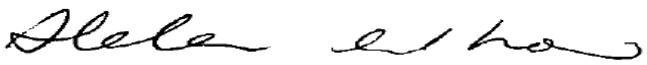
Declaration of the Candidate

I declare that I am the author of this thesis and unless otherwise stated, all references cited have been consulted by the candidate. The thesis is my own work, and has not been previously submitted for a higher degree.

Dr Harshal Deshmukh

Declaration of the Supervisor

I certify that Harshal Deshmukh has completed the equivalent of nine terms of experimental research and that he has fulfilled the conditions of the University of Dundee so that he is qualified to submit this thesis in application for the degree of Doctor of Philosophy.



Prof Helen Colhoun

Acknowledgements

I acknowledge the support of the Health Informatics Centre, University of Dundee for managing and supplying the anonymised data and NHS Tayside, the original data owner. I am grateful to all the participants who took part in the CARDS clinical trial and the Go-DARTS study, to the general practitioners, to the Scottish School of Primary Care for their help in recruiting the participants, and to the whole team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. I would like to thank all the groups that submitted their own data, which contributed to work presented in this thesis.

Part of this work was funded by Pfizer and by IMI SUMMIT study, under the EU Framework Programme 7 funding stream. The Wellcome Trust provides support for Wellcome Trust United Kingdom Type 2 Diabetes Case Control Collection (Go-DARTS) and the Scottish Health Informatics Programme.

I would like to thank my supervisor, Prof Helen Colhoun for her guidance and support throughout my PhD. I would like to thank Prof Paul McKeigue for his guidance on statistical and computational methods used during the PhD. I would like to thank Prof Ewan Pearson, Prof Colin Palmer, and Dr Helen Looker for their academic and personal support during my PhD. I would also like to thank my parents (Dr Arun and Mrs Seema Deshmukh, brother and sister in law (Vishwanth and Mrs Dhanashri Deshmukh), and wife (Mrs. Ketki Deshmukh) for their support through this process. Finally, I would like to thank H.H Sri Sri Ravishankar, and Shri Sant Shankar Maharaj (Pimpalkutha) who initiated me into meditation and have provided spiritual guidance through this journey.

Publications during PhD

1. **Harshal A. Deshmukh**, Jenny M. Wadén, Emma Ahlqvist , Barry I. Hudson4, Harald Grallert, Niina Sandholm, Natalie Van Zuydam, N. William Rayner, Kaixin Zhou, Paul Durrington, Helen Looker, Barbara Thorand, Barbara Nitz, Mark McCarthy , Leif Groop , Per-Henrik Groop, Colin Palmer, Paul McKeigue, Graham A Hitman, Helen M. Colhoun**: Genome-wide association Study identifies novel variants in the 3- UTR of ITGA1 and HLA-DQA1 associated with soluble receptor for advanced glycation endproducts (sRAGE) levels**.(Submitted Circulation Genetics)

2. **Harshal A Deshmukh**, Helen M Colhoun, Toby Johnson, Paul M McKeigue, D John Betteridge, Paul N Durrington, John H Fuller, Shona Livingstone, Valentine Charlton-Menys, Andrew Neil, Neil Poulter, Peter Sever, Denis C Shields, Alice V Stanton, Aurobindo Chatterjee, Craig Hyde, Roberto A Calle, David A Demicco, Stella Trompet, Iris Postmus, Ian Ford, J Wouter Jukema, Mark Caulfield, Graham A Hitman (2012) **Genome-wide association study of genetic determinants of LDL-c response to atorvastatin therapy: importance of Lp(a)**. J Lipid Res 53: 5. 1000-1011 May 2011

3. **Deshmukh HA**, Palmer CN, Morris AD, Colhoun HM ; **Investigation of known estimated glomerular filtration rate loci in patients with Type 2 diabetes** *Diabet Med*. 2013 Apr 16. doi: 10.1111/dme.12211.

4. Meng W1, **Deshmukh HA**, van Zuydam NR, Liu Y, Donnelly LA, Zhou K; Wellcome Trust Case Control Consortium 2 (WTCCC2); Surrogate Markers for Micro- and Macro-Vascular Hard Endpoints for Innovative Diabetes Tools (SUMMIT) Study Group, Morris AD, Colhoun HM, Palmer CN, Smith BH. **A genome-wide association study suggests an association of Chr8p21.3 (GFRA2) with diabetic neuropathic pain Eur J Pain**. 2014 Jun 26. doi: 10.1002/ejp.560. [Epub ahead of print]

5. Stella Trompet, Iris Trompet, **Harshal Deshmukh**.....Wouter Jukema, Mark Caufield; **Pharmacogenetic Meta-Analysis of GWAS of LDL Response to Statins** *Nature Communications*(in press)

6. S. Alkayyali, M. Lajer, **H. Deshmukh**, E. Ahlqvist, H. Colhoun, B. Isomaa,5, P. Rossing , L. Groop, V. Lyssenko**: Common variant in the HMGA2 gene increases susceptibility to nephropathy in patients with type 2 diabetes** *Diabetologia* 2012

7. Zhou K, Donnelly L, Yang J, Li M, **Deshmukh H**, Van Zuydam N, Ahlqvist E; The Wellcome Trust Case Control Consortium 2, Spencer CC, Groop L, Morris AD, Colhoun HM, Sham PC, McCarthy MI, Palmer CN, Pearson ER: **Heritability of variation in glycaemic response to metformin: a genome-wide complex trait analysis**: Lancet Diabetes Endocrinol. 2014

8. N van Leeuwen, G Nijpels, M L Becker, **H Deshmukh**, K Zhou, B H C Stricker, A G Uitterlinden, A Hofman, E van 't Riet, C N A Palmer, B Guigas, P E Slagboom, P Durrington, R A Calle, A Neil, G Hitman, S J Livingstone, H Colhoun, R R Holman, M I McCarthy, J M Dekker, L M 't Hart, E R Pearson (2012) **A gene variant near ATM is significantly associated with metformin treatment**

response in type 2 diabetes: a replication and meta-analysis of five cohorts. Diabetologia 55: 7. 1971-1977 Jul

9. Helen M Colhoun, D John Betteridge, Paul Durrington, Graham Hitman, Andrew Neil, Shona Livingstone, Valentine Charlton-Menys, Weihang Bao, David A Demicco, Gregory M Preston, **Harshal Deshmukh**, Kathryn Tan, John H Fuller (2011) **Total soluble and endogenous secretory receptor for advanced glycation end products as predictive biomarkers of coronary heart disease risk in patients with type 2 diabetes: an analysis from the CARDS trial**. Diabetes 60: 9. 2379-2385 Sep
10. Cristian Pattaro, Anna Köttgen, Alexander Teumer, Maija Garnaas, Carsten A Böger, Christian Fuchsberger, Matthias Olden, Ming-Huei Chen, Adrienne Tin, Daniel Taliun, Man Li, Xiaoyi Gao, Mathias Gorski, Qiong Yang, Claudia Hundertmark, Meredith C Foster, Conall M O'Seaghdha, Nicole Glazer, Aaron Isaacs, Ching-Ti Liu, Albert V Smith, Jeffrey R O'Connell, Maksim Struchalin, Toshiko Tanaka, Guo Li, Andrew D Johnson, Hinco J Gierman, Mary Feitosa, Shih-Jen Hwang, Elizabeth J Atkinson, Kurt Lohman, Marilyn C Cornelis, Åsa Johansson, Anke Tönjes, Abbas Dehghan, Vincent Chouraki, Elizabeth G Holliday, Rossella Sorice, Zoltan Kutalik, Terho Lehtimäki, Tõnu Esko, **Harshal Deshmukh**,et al…(2012) **Genome-wide association and functional follow-up reveals new loci for kidney function. PLoS Genet** 8: 3. 03

11. Daniel I Chasman, Christian Fuchsberger, Cristian Pattaro, Alexander Teumer, Carsten A Böger, Karlhans Endlich, Matthias Olden, Ming-Huei Chen, Adrienne Tin, Daniel Taliun, Man Li, Xiaoyi Gao, Mathias Gorski, Qiong Yang, Claudia Hundertmark, Meredith C Foster, Conall M O'Seaghdha, Nicole Glazer, Aaron Isaacs, Ching-Ti Liu, Albert V Smith, Jeffrey R O'Connell, Maksim Struchalin, Toshiko Tanaka, Guo Li, Andrew D Johnson, Hinco J Gierman, Mary F Feitosa, Shih-Jen Hwang, Elizabeth J Atkinson, Kurt Lohman, Marilyn C Cornelis, Asa Johansson, Anke Tönjes, Abbas Dehghan, Jean-Charles Lambert, Elizabeth G Holliday, Rossella Sorice, Zoltan Kutalik, Terho Lehtimäki, Tõnu Esko, **Harshal Deshmukh** et.al.. (2012) **Integration of Genome-Wide Association Studies with Biological Knowledge Identifies Six Novel Genes Related to Kidney Function**. Hum Mol Genet Sep 2012

12. Sandholm N, Salem RM, McKnight AJ, Brennan EP, Forsblom C, Isakova T, McKay GJ, Williams WW, Sadlier DM, Mäkinen VP, Swan EJ, Palmer C, Boright AP, Ahlqvist E, Deshmukh HA et.al: **New susceptibility Loci associated with kidney disease in type 1 diabetes PLoS Genet**. 2012 Sep;8(9):e1002921

Thesis Summary

Introduction

Diabetic kidney disease (DKD) is the leading cause of end-stage renal disease (ESRD), present in approximately 25%–40% of patients with long-standing diabetes and conferring additional risk of cardiovascular disease and mortality. Variations in the clinical presentations of DKD, heritability estimates from family-based studies and, more recently, the results from Genome-wide Association Studies (GWAS) demonstrate a heritable component of DKD. However, as is the case with the most of complex disorders, identifying causal genetic variants contributing to DKD has proven difficult. An important step in identifying variants associated with DKD in diabetes will involve integration of patient populations across multiple DKD cohorts, investigating rarer variants and by addressing the heterogeneity in DKD disease phenotypes in diabetes.

Methods

In this thesis, I reviewed the existing literature in genetic epidemiology in diabetic kidney disease. I then estimate chip-based heritability of DKD sub-phenotypes and replicated the association of known SNPS associated with renal function and upstream risk factors for diabetic kidney disease (BP, HbA1c) in patients with Type 2 Diabetes. I performed first GWAS for soluble receptor for advanced glycation products (sRAGE) a biomarker implicated in the pathogenesis of DKD. Finally, I performed GWAS for various DKD phenotypes on Type 1 Diabetes cohort (EURODIAB) and Type 2 Diabetes cohort (Go-DARTS) and helped with joint meta-analysis with DKD cohorts in SUMMIT consortium investigating genetic determinants of DKD.

Results

First, I showed that some DKD sub-phenotypes (like macro-albuminuria and ESRD) might be more heritable than others are and demonstrate that usefulness of estimation of chip-based heritability for complex trait by GCTA can be limited in the absence of large sample sizes. Second, I investigated the known genes for renal function (eGFR) and upstream risk factors for diabetic kidney disease (BP, HbA1c) in patients with Type 2 diabetes and showed that cumulative genetic risk for BP and HbA1c is associated with DKD. Third, I replicated the association of known loci associated with eGFR (*UMOD GCKR and SHROOM3*) in patients with Type 2 diabetes and showed that albuminuria affects the association of these variants with renal function. Fourth, I conducted a GWAS for sRAGE, an important biomarker associated with DKD, and identified novel variants in *ITGA1* and *HLA-DQA1* associated with circulating sRAGE levels. Finally, I performed GWAS for various DKD sub-phenotypes, and assisted in GWAS meta-analysis with SUMMIT consortium and identified potential novel genetic determinants for diabetic kidney diseases.

Conclusion

In conclusion this thesis has shown that a) estimation of chip based heritability of various DKD sub-phenotypes using GCTA has limited utility and requires GWAS studies with extremely large sample sizes b) the genetic determinants of renal function (eGFR) can interact with albuminuria in patients with T2D c) there are yet unidentified genetic markers associated with DKD and have identified potentially novel genetic markers associated with sRAGE (an important biomarker for DKD) and DKD itself which can be investigated in future studies for their reproducibility and functional consequences.

Background and Literature Review

I. Introduction

There is an ongoing trend of a rapid increase in the incidence of diabetes mellitus, especially the non-insulin dependent form. By the end of the second millennium, 150 million cases were recorded worldwide, while the estimations predicted doubling the number by the year 2030[1]. Numerous chronic complications accompany the disease; among them micro- as well as macro-vascular prevail, affecting small and large blood vessels such as diabetic kidney disease (DKD), retinopathy, cardiovascular disease, and lower-extremity arterial disease. Here we review the existing literature on pathophysiology, heritability, and genetic epidemiology of DKD.

II. Definition of Diabetic Nephropathy

Diabetic nephropathy (DN) is a clinical diagnosis, which is historically based on the findings of persistent urinary protein excretion in the absence of any known cause of renal disease. This definition was initially confined to those who are now considered to have macro-albuminuria (Table 1). The development of more sensitive assays specific for albumin has since led to the detection of smaller increases, termed micro-albuminuria or "incipient nephropathy". The lower limit to confer micro-albuminuria diagnosis is an albumin excretion rate (AER) of 20 µg/min, which is equivalent to 30 mg/24 h or an ACR of 30 mg/g[2] (Table1). These definitions are clinically relevant as individuals with macro-albuminuria show a progressive decrease in renal function (GFR) associated with an increase in systemic blood

pressure, whereas those with micro-albuminuria are considered to have stable kidney function but, are at high risk of subsequent development of macro-albuminuria and kidney failure[3] . Table 1 shows the cut-off values according to the National Kidney Foundation Kidney Disease Outcomes Quality Initiative (*NKF-KDOQI*).

Table 1 Definitions of the abnormalities of albumin excretion

Category	Spot collection mg/g creatinine	24 hour collection(mg/24 hour	Timed collection ug/min
Normoalbumiuria	<30	<30	<20
Microalbuminuria	30-300	30-300	20-200
Macroalbuminuria	>300	>300	>200

Because of the variability in the UAE, two or three specimens collected within 3-6 months period should be abnormal before considering the patient to have crossed one of the diagnostic thresholds. It is not uncommon to get false positive results because of infections, fever, congestive heart failure, hypertension, hyperglycaemia, etc. and both true progression and regression can occur. Notably, regression of micro-albuminuria in Type 1 Diabetes is quite frequent. Krolewski AS et.al showed that six-year cumulative incidence of regression of micro-albuminuria can be upto 58 percent in patients with Type 2 Diabetes[4]. This study showed that lower glycosylated haemoglobin , lower systolic blood pressure , and lower levels of both cholesterol and triglycerides were independently associated with the regression of micro-albuminuria.

III. Concept of Chronic Kidney Disease (CKD) in Diabetes

Glomerular filtration rate (GFR) can be a useful indicator of renal function in diabetes, and is known to decline in the natural history of diabetic nephropathy following onset of macro-albuminuria[5]. Based on the eGFR levels CKD has been divided into following stages[6].

Stage 1: Slightly diminished function; kidney damage with normal or relatively high GFR (≥90 mL/min/1.73 m2)

Stage 2: Mild reduction in GFR (60–89 mL/min/1.73 m2) with kidney damage

Stage 3: Moderate reduction in GFR (30–59 mL/min/1.73 m2). British guidelines distinguish between stage 3A (GFR 45–59) and stage 3B (GFR 30–44) for purposes of screening and referral

Stage 4: Severe reduction in GFR (15–29 mL/min/1.73 m2)

Stage 5: Established kidney failure (GFR <15 mL/min/1.73 m2, permanent renal replacement therapy (RRT), or end stage renal disease (ESRD)

Given the difference in the risk factors for GFR decline and albuminuria, it is not surprising to see that eGFR decline can precede the onset of albuminuria in patients with diabetes[7]. Hence, eGFR alone cannot predict or diagnose diabetic kidney disease per se. Indirect estimation of GFR using various formulae is described in the methods section, below.

IV. Concept of Diabetic Kidney Disease

Diabetic kidney disease (DKD) is a term used to annotate kidney disease in diabetes using both eGFR and albuminuria as eGFR and albuminuria alone cannot reflect renal damage of attributed to diabetes. The likelihood of diabetic kidney disease (DKD) according to Staging by eGFR and albuminuria levels is summarised in Table 2 (NKF KDOQI guidelines)[6]. Those with eGFR <60 mls/min/1.73m2 or below with normo-albuminuria and those with CKD Stage 4 or above with micro-albuminuria are unlikely to have DKD highlighting the point that eGFR alone cannot be used in the diagnosis of DN. Although these combinations of eGFR and albuminuria status, and other unusual clinical presentations, warrant a biopsy to make a definitive diagnosis of DKD[8] mostly DKD is a clinical diagnosis and biopsy is not routinely carried out. This of course has implications for genetic studies because the phenotype is not a clean one. Especially in type 2 diabetes up to 50% of those with declining eGFR have non diabetic causes for their chronic kidney disease (CKD)[9] and albuminuria is not specific for having true diabetic glomeruloscerlosis on biopsy it just increases the probability of this[10].

V. Role of Retinopathy in Diagnosis of DKD

The concomitant presence of retinopathy is partly helpful in discriminating kidney pathology in patients with Type 2 diabetes[11]. The presence of retinopathy in patients with Type 2 diabetes and macro-albuminuria is strongly suggestive of DKD, and its absence in micro-albuminuria suggests non-DKDs though it is not specific for DKD. The sensitivity (ability of test to detect true positive), specificity (ability of test to detect true negative), positive predictive value (proportions of positive results in statistics and diagnostic tests that are true positive), and negative predictive value

(proportions of negative results in statistics and diagnostic tests that are true negative) of retinopathy in the diagnosis of DKD is well documented. In those with macro-albuminuria, the positive predictive value (PPV) of retinopathy for typical diabetic glomerulopathy ranges from 67% to 100%. However, the negative predictive value (NPV) had a broader range of 20% to 84% (sensitivities between 26% and 85% and specificities of 13% to 100%)[12]. For micro-albuminuria, PPVs of retinopathy were lower at around 45% but NPVs were close to 100%, giving sensitivities of 100% and specificities of 46% to 62%.

Table 2: Likelihood of DKD according to Staging by eGFR and albuminuria level[6]

GFR	CKD Stage	Normbuminuria	Microalbumniuria	Macroalbuminuria
>60	1+2	At Risk	Possible DKD	DKD
30-60	3	Unlikely DKD	Possible DKD	DKD
<30	4+5	Unlikely DKD	Unlikely DKD	DKD

VI. Natural History of Diabetic Nephropathy and Nephropathy Phenotypes

In Type 1 Diabetes, the earliest sign of DKD is often micro-albuminuria, which appears within 5-15 years of duration of diabetes, and the incidence of micro-albuminuria over the lifetime is about 50%[13]. Recent studies have suggested that about 40% patients with micro-albuminuria will progress to macro-albuminuria[14]. If untreated, with a progressive decline in GFR, almost all the patients with macro-albuminuria will develop end-stage renal disease ESRD and those with Stage 3 CKD die prematurely of cardiovascular events[15].

The development of DKD in T2D follows a similar course as in T1D patients but the lifetime prevalence of proteinuria is lower and varies according to ethnicity[16].

Studies have shown that DKD is more prevalent and develops more rapidly in African-Americans, Asians, and Native-Americans than in Caucasians. The reported prevalence of DKD in Pima Indians is about 50% within 15-year follow-up of diabetes[17], while it is about 30% in African Americans[18]. In the Caucasians populations on the other hand, the prevalence is about 20% in European Americans (WESDR)[19] and 5% in a UK population (UKPDS)[5] over a 10-year follow-up period. The natural progression of diabetic nephropathy consists of successive, coinciding pathologic and hemodynamic changes, which correspond to different stages of albumin excretion (Table 3). The initial phase is characterised by renal hypertrophy, an increase in GFR, and blood pressure and normo-albuminuria. Those who are resistant to subsequent changes (either because of protective genetic effects or environmental effects) tend to stay in this stage for a longer period (15 years or more) than others and are usually used as controls in genetic studies for DN. The next phase is characterised by glomerular basement thickening, increase in GFR, and blood pressure, clinically manifesting as micro albuminuria. Subsequently, there is meningeal expansion and onset of fibrosis with decline in renal function, clinically manifesting as macro-albuminuria. Finally progressive renal fibrosis replaces the normal renal tissue resulting in minimal renal function (GFR<15) leading to elevated serum urea and creatinine resulting in end-stage renal disease (ESRD). The genetic determinants of the pathological and hemodynamic changes in each of these phases are not definitively characterised. A classification of DKD based on various stages of Albuminuria and eGFR can help in identifying genetic determinants specific to the spectrum of underlying pathologies in DKD. For example, a genetic association study involving patients with micro-albuminuria and normo-albuminuria might identify the genetic variants associated with basement membrane thickening,

whereas a study involving patients with macro-albuminuria and normo-albuminuria can identify the genetic variants associated with mesangial expansion, fibrosis and a study involving patients with ESRD can uncover the genetic determinants of renal fibrosis.

Table 3 Nephropathy phenotypes and corresponding hemodynamic and pathologic changes

Phenotypes	Diabetes and Normo-albuminuria	Micro-albuminuria	Macro-albuminuria	ESRD
Hemodynamic Changes	↑ GFR and ↑BP	↑ GFR and ↑BP	↓ GFR and ↑BP	↓ ↓GFR, ↑BP
Predominant Pathologic Changes	Renal Hypertrophy	Basement membrane thickening	Basement membrane thickening , Mesangial expansion and onset of global fibrosis	Progressive global fibrosis

VII. Current Understanding of Key Aspects of Histology and Pathogenesis of DKD including some Key Pathways

Histopathology of Diabetic Kidney Disease

The morphologic lesions in diabetic nephropathy are predominantly seen in the glomeruli, with an observed thickening of the glomerular basement membrane (GBM) and mesangial expansion early in the disease. In Type 1 diabetes, these changes occur as early as 1.5 to 2.5 years after the onset of T1DM[20]. Later in the course of the disease the podocytes, renal tubules, interstitium, and arterioles undergo substantial changes. The hallmark of DKD histology is the 'Kimmelstiel-Wilson nodules' where there is diffuse mesangial expansion, (also termed diffuse diabetic glomerulosclerosis); the nodular lesions consist of areas of marked

mesangial expansion forming large round fibrillar mesangial zones with palisading of mesangial nuclei around the periphery of the nodule and compression of the associated glomerular capillaries. Mesangial expansion, predominantly due to an increase in mesangial matrix, can be detected as early as 5–7 years after the onset of diabetes[21]. When the mesangium expands, it restricts and distorts glomerular capillaries and diminishes the available capillary filtration surface resulting in decline in renal function and a highly significant inverse correlation between mesangial expansion and GFR, AER and blood pressure[22]. GBM thickening on the other hand is closely related to AER and less so to GFR or hypertension, suggesting that this lesion is a closer surrogate to the pathogenesis of albuminuria[22].

The interest in role of podocytes in DKD is increasing in recent years, although there are no specific evidences. Podocyte detachment from GBM, from apoptosis, necrosis, or loss of adhesive interaction, may play a role in the pathogenesis of proteinuria. It has been shown that proteinuria in glomerular disorders is associated with foot process effacement, flattening and retraction[23].

Physiological Changes in DKD

DKD has several distinctive phases of development. Before the onset of clinical changes and appearance of albuminuria, functional changes occur in the kidneys. The process starts with hyper-filtration and hyper-perfusion of the glomeruli. Hyper-filtration is typically defined as GFR between 125 mL/min to 140 mL/min per 1.73 m^2, or greater than two standard deviations above the mean GFR in normal healthy individuals[24]. Hyper-filtration is observed in 25%-75% of patients with T1D and about 5-40% patients with T2D. The degree of hyper-filtration appears to be related to impairment of fasting glucose and duration of diabetes in the

general population[25]. For example, in a study of a Pima Indian population[26], those with a normal glucose tolerance test had mean eGFR of 123 mL/min, and those with impaired glucose tolerance test had mean eGFR of 135 mL/min. In the same study, individuals with newly diagnosed diabetes mean eGFR was 143 mL/min and, those with overt diabetes for more than five years and either normal albumin excretion or micro-albuminuria, the mean GFR was 153 mL/min.

Although the pathogenesis of diabetic hyper-filtration is not completely understood, several hypothesis implicating glomerular hemodynamic and tubular factors have been proposed to explain the mechanisms that are responsible for hyper-filtration[27,28].

There is a good experimental evidence[29] to suggest that hyperglycaemia increases the production and availability of vasoactive mediators that regulate glomerular arteriolar tone, such as, nitric oxide (NO) system, cyclooxygenase 2 (COX2)-derived prostanoids, the renin angiotensin system (RAS), protein kinase C (PKC) and endothelin (ET). These vasoactive substances change pre-glomerular (afferent) and post-glomerular (efferent) arteriolar tone, such that there is a decreased resistance at the afferent arteriole as compared to the efferent arteriole[29]. This causes an increase in the blood flow through the glomerulus, which eventually leads to leakage of albumin from the glomerular capillaries, a thickening of the glomerular basement membrane and injury to the podocytes. The mechanical strain resulting from these hemodynamic changes can induce localised release of cytokines (TGF-β1) and growth factors, which cause increased synthesis of collagen and fibrosis[30] .

The tubular hypothesis suggests that hyper-filtration is initiated by increased sodium reabsorption in the proximal tubule, which is mediated by the sodium-glucose cotransporter-2 (SGLT2)[31]. This increase in proximal reabsorption reduces sodium

signalling to the macula densa, which sense a decline in effective circulating volume and renal perfusion. The reduction in renal perfusion reduces adenosine in the juxtaglomerular apparatus of the kidneys leading to the dilatation of the afferent arteriole, hyper-perfusion, and hyper-filtration[32]. Animal studies have supported the tubular hypothesis by showing that SGLT2 inhibition decreases hyper-filtration and diminishes the histological evidence of diabetic nephropathy[33]. There are no human studies looking at the effects of these agents on eGFR however, they have been shown to be associated with lower blood sugar, blood pressure and lower weight[34].

Role of Cytokines and Vascular Growth Factors in DKD

Activation of cytokines (TGFβ-1) and vascular growth factors such as VEGF might play a role in the matrix accumulation that arises in diabetic nephropathy[35,36].

There is a strong evidence showing that TGFβ-1contributes to the cellular hypertrophy and increased synthesis of collagen, both of which occur in diabetic nephropathy[37,38]. TGFβ-1 levels are increased in the glomeruli of rats with streptozotocin-induced diabetes, and antibody to TGFβ-1 prevents renal changes of diabetic nephropathy in these animals[39]. In addition, connective tissue growth factor and heat shock proteins, which are encoded by TGF-1-inducible genes, have fibrogenic effects on the kidneys of patients with diabetes[40].

There is some evidence to that VEGF increases permeability of the glomerular filtration barrier to proteins; however, surprisingly the levels of this growth factor can be low in rats with diabetes[41]. Further evidence to support a pathogenic role for VEGF in diabetic nephropathy comes from the fact that VEGF blockade improves albuminuria in an experimental model of the disorder[35,36]. However, some studies refute a causative role for high VEGF levels in diabetic nephropathy and suggest that

VEGF plays important role in mesangial cell development. Eremina et al[42] demonstrated in a mouse model that VEGF is produced by podocytes and is crucial for glomerular endothelial cell survival and differentiation as well as for mesangial cell development and differentiation. Gene expression of VEGF is decreased in humans with diabetic nephropathy, however it is not known if this is due to the podocyte loss in DKD. In summary, the role of VEGF in the pathophysiology of nephropathy is largely unclear.

Inflammatory cytokines also contribute to the development and progression of diabetic nephropathy, specifically interleukin 1 (IL-1), IL-6 and IL-18, and tumour necrosis factor and levels of these cytokines in serum and urine correlate with progression of nephropathy, as indicated by increased urinary albumin excretion[43].

Role of AGE-RAGE Axis in DKD

There is growing body of evidence suggesting that AGE-RAGE axis is involved in the pathogenesis of DKD[44]. Among various types of AGE receptors, RAGE is a signal transducing receptor for AGEs that could mediate the inflammatory reactions evoked by AGEs[45]. In humans, RAGE expression is enhanced in podocytes and mesangial cells in diabetic patients with nephropathy[46]. Animal studies show that RAGE-overexpressing diabetic mice have progressive glomerulosclerosis with renal dysfunction, compared with diabetic littermates lacking the RAGE transgene[47]. Studies have reported that diabetic homozygous RAGE null mice fail to develop mesangial matrix expansion or thickening of the glomerular basement membrane[44]. It has also been shown that streptozotocin-induced diabetic mice develop renal changes seen in human diabetic nephropathy such as glomerular hypertrophy, glomerular basement membrane thickening, mesangial matrix expansion, connective

tissue growth factor (CTGF) overexpression, and NFκB activation - all of which are blocked by the administration of neutralizing antibody raised against RAGE[48]. The AGE-RAGE interaction can also induce sustained activation of NFκB because of increased levels of de novo synthesized NFκBp65 overriding endogenous negative feedback mechanisms and thus might contribute to the persistent damage to diabetic kidney[49].

Role of Oxidative Stress DKD

Hyperglycaemia induces oxidative stress, even before diabetes becomes clinically apparent. Concentrations of markers of DNA damage induced by reactive oxygen species are higher in patients with more-severe nephropathy (i.e. proteinuria versus micro-albuminuria). Furthermore, histological analysis of human kidney biopsy specimens has detected products of glycooxidation (combined products of glycation and protein oxidation) and lipooxidation in the mesangial matrix and glomeruli, whereas these lesions are much less common in specimens from individuals without diabetes[50,51]. However, the mechanisms by which the free oxygen radicals are involved in the pathogenesis of DKD are not well understood.

Role of Lipids in Pathogenesis of DKD

Progressive renal failure associated with proteinuria, is accompanied by abnormalities of lipoprotein transport[52]. Typically, in DKD there are increased serum levels of triglycerides ,cholesterols, VLDL, apoB and pre-β HDL, and low levels of HDL and of apoA[53]. The mechanisms by which elevated lipids can progress the onset of DKD are unclear. However, there is evidence that circulating lipids bind to and become trapped by extracellular matrix molecules , where they undergo

oxidation which increases the formation of reactive oxygen species and reduces the actions of endothelium-derived vasodilators/growth inhibitors[54]. This increase in reactive oxygen species and decrease in vasodilators can have significant vascular and renal pathophysiologic effects. Small lipids derived from arachidonic acid have also been implicated in the pathogenesis of DKD. Cyclo-oxygenase 2 breaks down arachidonic acid into several different prostanoids[55] and in a rat model of streptozotocin-induced diabetes, levels of inflammatory prostanoids, such as prostaglandins E2 and I2, were raised. However, a more detailed characterisation is needed of how the production of prostanoids affects the pathogenesis of DKD in humans.

VIII. Heritability of diabetic nephropathy

Heritability of DKD has been established by demonstration of familial clustering of the disease and estimations of heritability of albuminuria and eGFR using family based studies.

Familial Clustering of Diabetic Nephropathy

Seaquist et.al first demonstrated familial clustering of DKD[56]. The study showed that out of the 29 diabetic siblings of probands with diabetic nephropathy, 24 (83%) had evidence of nephropathy including 12 with end-stage renal disease while evidence of nephropathy was found in 2 of the 12 diabetic siblings of the probands without nephropathy (17%). A few years later in 1992 a study in Denmark[57] looked at of 619 patients with IDDM; the study identified 20 patients with and 29 without nephropathy as having diabetic siblings. Diabetic nephropathy was found in

7 out of 21 siblings to patients with nephropathy and 3 out of 30 siblings to normo-albuminuria patients. There were no significant differences between the two groups of siblings with respect to age, diabetes duration, sex distribution, blood pressure, or glycosylated haemoglobin A1c- levels. A study in Pima Indians, a homogeneous population with a high occurrence of NIDDM showed that the families in which two successive generations had Type 2 diabetes the likelihood of the offspring developing overt nephropathy was 14% if no parent had proteinuria, 23% if one parent had proteinuria and 46% if both parents had proteinuria[58]. Subsequently, the familial aggregation of DKD was also observed in India[59], South America,[60] and Asia[61]. In spite of the repeated confirmation, there was always concern that all these single centre reports of familial aggregation could be biased and the observed clustering could have been a result of shared environment, for example, lack of access to health care or an unrecognised environmental risk factor. Finally a large multi-centre analysis in 2005[62] involving 26,000 incident dialysis patients in 450 dialysis clinics across the United States were screened for family history of ESRD. After removal of individuals with known genetic or urologic disorder, the study found that 32% of women and 27% of men with African ancestry and a 15% and 12% of European American women and men respectively reported having a close relative with ESRD. In a small subset of the same dataset[63], 66 unrelated index African American cases with overt Type 2 DN/ESRD, 132 of their diabetic sibs, and 13 of their non-diabetic sibs, more than 60% of index cases had at least one diabetic sibling with overt proteinuria. The findings of this large multi-centre study confirmed the earlier single centre reports across the globe and firmly established familial aggregation of DKD.

Heritability Estimates for Diabetic Nephropathy

Heritability estimates for diabetic nephropathy have been reported for both albuminuria (urine albumin-creatinine ratio (ACR)) and renal function (GFR), confirming genetic contributions to the pathogenesis of DKD.

Heritability measures the fraction of phenotype variability that is because of the genetic variation. Most of common complex diseases have both environmental and inherited effects and the purpose heritability is measure of genetic component of the disease [64]

"H^2" is the broad-sense heritability, which is the sum of all the sources of genetic variations that can be attributed to the disease. "h^2" on the other hand is a capture of whole genome additive genetic variations as measured by a microarray, which can be attributed to disease and is also known as "chip based" or narrow sense heritability.

Heritability has been estimated for both albuminuria and eGFR in patients with T1D and T2D. A heritability (H^2=1) suggests a Mendelian disorder while and heritability (H^2 <1) suggests some contribution from environment. Estimates of heritability for albuminuria varied from 0.30 to 0.44 in Finnish[65], New England[66], and south-eastern US[67] families enriched for members with Type 2 diabetes, and the estimate was slightly larger in (0.49) in Hypertension Genetic Epidemiology Network (HyperGEN) families enriched for multiple siblings with hypertension[68]. The heritability estimates for GFR on the other hand varied from 0.36 to 0.75[67,69,70] in the populations with European ancestry.

Differences in the heritability of albuminuria and eGFR suggest presence of distinct genetic loci, and pathogenic mechanisms affecting renal vascular permeability and renal function in general. It is known that kidney diseases characterized by

albuminuria, such as diabetic nephropathy can have ultrafiltration and high eGFR in the early stage of disease while those characterised by reduced renal function like hypertensive kidney disease may manifest with normo-albuminuria because of the reduced renal efficiency[71,72]. Differences in the heritability estimates for albuminuria and eGFR, and the fact that both albuminuria and eGFR influence each other during the course of diabetes indicate the need to adjust for albuminuria while estimating the loci for eGFR and vice-versa.

IX. Animal Models for Diabetic Kidney Disease

At present, there are no reliable animal models that completely mimic human DKD. Chemical agents such as streptozotocin (STZ) can selectively damage the insulin-producing beta cells in the pancreas resulting in hyperglycaemia and is considered an important tool for developing animal models of diabetic complications[73]. While this model induces kidney hypertrophy and mesangial expansion, it does not progress to more advanced renal disease as seen in humans (loss of glomerular filtration, overt proteinuria, advanced structural lesions and tubulointerstitial fibrosis)[73]. Mouse models of endothelial dysfunction, with a targeted mutation in the *NOS3* gene encoding endothelial nitric oxide synthase are one of the more robust models of advanced renal disease in diabetes; these animals exhibit both decline in glomerular filtration and tubulointerstitial fibrosis[74]. A recently developed mouse model with a leptin receptor mutation develop Type 2 diabetes, hypertension, obesity and proteinuria, reduced glomerular filtration, mesangial matrix expansion and podocyte loss[75], thereby mimicking diabetic kidney disease. Although no single-animal model exists, that shows exact pathophysiological features of established DKD Type 1 and

Type 2 diabetes, these have provided valuable information regarding many aspects of DN including pathophysiology and putative roles of implicated genes. Having better animal models for DKD are important to test the hypothesis generated by GWAS studies for DKD and to test treatment strategies based on those hypothesis.

X. Candidate Gene Studies, Sib pair Studies and GWAS Studies for Kidney Disease

Literature Review Candidate Gene Studies for Diabetic Nephropathy and CKD in Diabetes

The candidate gene approach to conducting genetic association studies focuses on associations between genetic variation within pre-specified genes of interest and phenotypes or disease states. Most often, selection of candidate genes for study is based on *a priori* knowledge of the gene's biological functional impact on the trait or disease in question. For DKD, the choice of the genes for study depends on knowledge concerning its actions in DKD pathophysiology such as those involving blood pressure control, severity of proteinuria, insulin resistance, lipid metabolism and other mechanisms implicated in the pathogenesis of diabetic nephropathy. There are about 209 genes, which have been investigated for their association with DKD using a candidate gene approach. Table 4 summarizes some of the commonly investigated genes associated with Diabetic Nephropathy and the review briefly discusses the top four biologically relevant and candidate genes for DKD.

Table 4: Top Candidate Genes for Diabetic Nephropathy

Gene	Number of Association Studies	Number of Meta-analysis	Involved KEGG pathways
ACE	56	4	Renin-angiotensin system
NOS3	24	5	Calcium signalling pathway, VEGF signalling pathway,
TGFB1	15	1	TGF-beta signalling pathway, Colorectal cancer, Cell Cycle
MTHFR	22	3	One carbon pool by folate, Methane metabolism
AGT	22	2	Renin-angiotensin system
APOE	21	2	Alzheimer's disease, Neurodegenerative diseases
AGTR1	19	2	Retroactive ligand-receptor interaction, Calcium signalling pathway
CNDP1	12	1	Histidine metabolism, beta-Alanine metabolism
PPARG	11	1	PPAR signalling pathway, Endometrial cancer
AKR1B1	11	2	Glycolipid metabolism, Pyruvate Metabolism Galactose metabolism
RAGE	10	0	-
ADIPOQ	9	1	Adipocytokine signalling pathway, PPAR signalling pathway, Type II diabetes mellitus
CCR5	9	1	Cytokine-cytokine receptor interaction
ELMO1	7	1	-

Angiotensin Converting Enzyme gene (ACE)

The angiotensin-converting enzyme (ACE), a potent vasoconstrictor, catalyses the conversion of angiotensin I to angiotensin II and inactivates bradykinin, a vasodilator; by proteolysis[76]. The angiotensin-converting enzyme gene (ACE) on chromosome 17q23 has repeatedly been evaluated for a role in DKD. In 1994, the

first evidence of involvement of ACE genotypes with DKD was reported by a French group[77]. The study compared 62 IDDM subjects with diabetic nephropathy with 62 diabetic controls with normal kidney function - with respect to the insertion/deletion polymorphism of the *ACE* gene (which predicts circulating ACE levels). This low powered study observed an imbalance of *ACE* genotype distribution, with a low proportion of subjects homozygotes for insertion at *ACE* polymorphism, was observed in IDDM subjects with diabetic nephropathy compared with their control subjects with diabetes without nephropathy. Subsequent to these observations, several reports contested these findings and reported no association of *ACE* gene with DKD[78-82]. Following these a first large-scale, prospective multi-centre study (17 centres in France and Belgium) on insulin-dependent diabetic subjects, GENEDIAB[83], looked at the association of insertion/deletion (I/D) polymorphism with severity of nephropathy and reported a positive association, chi2 for trend 5.135, P = 0.023; adjusted odds ratio attributable to the D allele 1.889 (95% CI 1.209-2.952, P = 0.0052). Given its biological functions and its relevance to the renal physiology, *ACE* seems to be a compelling candidate for DKD susceptibility; however, several non-replications and lack of evidence from GWAS studies suggest that it might contribute to small proportion of variability of DKD in Patients with Type 2 diabetes.

Nitric Oxide Synthase 3 (NOS3) Gene

Nitric Oxide has been reported to be an important regulator of renal hemodynamic in the vascular endothelium[84-86] and Nitric Oxide production can be influenced by polymorphisms of the NOS gene. Polymorphisms, which can decrease Nitric Oxide production, can increase in arterial pressure and thus can be associated in the

pathways of renal disease in DM. The first report of association of the Glu298Asp mutation of the NOS3 gene was shown in a study[85] in Japanese population consisting of 159 patients with all cause-ESRD undergoing maintenance haemodialysis and 270 genetically healthy control subjects. This showed increased frequency of Glu298Asp(rs1799983) or eNOS-894(G894T) mutation in patients with ESRD (22.0%), non-diabetic renal diseases (22.5%) and diabetic renal disease (20.8%) as compared to compared with controls. Since then, three polymorphisms in the NOS3 genes, eNOS-4b/a, eNOS-894(rs1799983) and eNOS-764(rs2070744) have been studied extensively for their association with DN. A recent study by Yanming He et al[87] conducted a meta-analysis of 18, 10 and 3 eligible studies eNOS-4b/a, eNOS-894(G894T) and eNOS-764(T786C) respectively. 3,793 patients (DKD) and 3,161 controls (diabetes without DKD) for 4b/a, 2,654 patients and 1,993 controls for eNOS-894 and 1,348 patients and 1,175 controls for eNOS-764 were included in the analysis. This study showed that the eNOS-4b/a was significantly associated in the overall meta-analysis with a stronger association in Asian population (OR 1.10 overall vs. OR=1.7 in Asian) and showed no association in the Caucasian subgroup. Similarly, the eNOS-894 polymorphism showed association only in the Asian population suggesting an ethnicity specific association of *NOS3* gene with DKD. The eNOS-764 showed a weak association in overall study population and did not have enough power to detect ethnicity specific association. A second meta-analysis with additional data for eNOS-894(G894T), (3,585 cases and 3,140 controls) showed that that the 894T was negatively associated with DKD in Caucasian populations of European origin but was positively associated with DKD in East Asian[88]. Taken together, results so far suggest that polymorphisms in *NOS3* gene can alter susceptibility to DKD in DM; however, this association could be ethnicity specific.

Apolipoprotein E (APOe)

APOE is the coding gene for apolipoprotein E. The genetic association studies for diabetic nephropathy have looked at three APOE alleles, ε2 ε3 and ε4 encoding three isoproteins E2 E3 and E4. A recent meta-analysis[89] of 17 studies in the Asian populations show that those with an APOε2 had almost double the risk of DKD as compared to those without it (OR=1.85, CL=1.49, 2.29). Studies which looked at ε4 alleles did not find an association of ε4 with DKD overall, however, a subset analysis in Chinese population showed an increased risk of DKD (OR=1.51, CL 1.11-2.06) and no association in the Korean and Japanese populations. A meta-analysis in European population[90], showed no association of the APO- ε2 variants with DN while no European study has looked at the effect of APO- ε4 on DKD. The differential effect of the APOe alleles on DN by ethnicity suggests either a possible genetic heterogeneity in the mechanisms through which the APOE affects DKD or presence of additional environmental factors (lifestyle, BMI) which interact with APOE to produce the DKD phenotypes.

TGFβ1 (Transforming growth factor beta 1)

TGFβ1, a fibrogenic cytokine with strong regulatory effects on renal cell hypertrophy and extracellular matrix accumulation is known to play a pivotal role in the initiation and progression of DKD in animal models[91]. Animal research has shown that mice over-expressing TGF-b1 were affected by progressive renal failure[92], and administration of anti-TGF-b1 antibody could inhibit glomerular hypertrophy and excessive extracellular matrix gene expression, and thus prevent the development of DKD[93]. Therefore, TGFβ1 is considered as a functional candidate for DKD. Answer: In diabetic nephropathy matrix accumulation in both glomeruli

and the interstitium correlates with both degree of renal insufficiency and proteinuria[94] . TGF-beta is potent fibrotic cytokine and therefor of considerable interest in DKD. Involvement of TGF-beta in T2D-DKD suggests that T2D-DKD is a pro-fibrotic chronic inflammatory condition. It also suggests that a better understanding of the activation of TGF-β signaling and its downstream regulators may provide new insights for the prevention of progressive diabetic nephropathy. A recent meta-analysis[95] of nine studies looked at association of TGFβ1 and DN in Type 1 and Type 2 diabetic nephropathy. This study showed significant association of TGF-b1 T869C polymorphism in both Asian and European population with DKD only in Type 2 diabetic subjects, but not in Type 1 diabetic patients. None of the candidate studies till date have shown an association of TGFβ1 with T1D –DKD[96]; however, a recent GWAS in patients with Type 1 diabetes suggests that that AFF3 gene can influences renal tubule fibrosis via the transforming growth factor-beta (TGF-β1) pathway[97]. TGFβ1, given its functional significance is a promising candidate for DN and further studies looking at the association of TGFβ1 with DN and its interaction with other candidate genes are warranted.

Candidate Gene Studies Summary

The limitations of genetic sequencing technology restricted the search for genetic variants, which influence DKD to association studies of 'likely genes' that are selected based on their pathways and knowledge of the pathophysiology of DN. This approaches although very appealing, has some inherent limitations: 1) Our understanding of the pathophysiology of DKD is limited and all the pathways, which can explain the predisposition, onset, and progression of DKD, are not precisely defined. Thus, this restricts the selection of 'likely genes' for candidate gene studies.

2) The candidate gene studies typically a have small sample size and very limited power to detect the variants, which cause modest increase in risk of DKD. It is possible to increase the power by meta-analysis (combing the published data on genetic variants across a number of studies and ethnic groups); however, these studies suffer from heterogeneity introduced by different study designs different phenotypes, genetic models and population stratification. A recent study by Mooyaart et al[90] combined data from 671 candidate gene association studies investigating genetics of DKD. In this study only three genetic variants met stringent criteria for significance, the ACE I/D APOE and AKRB1 in a random effects model, highlighting the limited power of these candidate gene meta-analyses. 3) Finally the non-replication of most of the genes identified by candidate gene studies (in the recent GWAS analysis) suggest possibility of publication bias, wherein marginally significant, falsely positive findings appear more in the literature whereas falsely negative studies are not published. With the advent of newer and cheaper sequencing technologies, candidate gene studies for DKD are performed to examine 'extremely compelling candidates' in novel pathways or for replication of genes already identified through GWAS.

Literature Review of Genome-wide Sib-pair studies for DKD

If a marker is significantly more common in the family members with DKD, it indicates presence of a susceptibility region on the chromosome associated with the phenotype of interest. For T1D usually both parents with affected offspring (trios) are included in the study, while in T2D the analysis involves sibling pairs since parents are usually unavailable at the time of recruitment. Over the last decade, several genome wide linkage scans looking to identify susceptibility to T1-DKD and

T2-DKD have been reported (Table 5). These studies have identified several chromosomal regions that appear to be associated with the development and progression of diabetic nephropathy. Follow-up studies of the linkage peaks thus identified have demonstrated associations of several plausible candidate genes with DN. For example, four independent genome-wide linkage studies have identified chromosome 3q as a major locus for diabetic nephropathy susceptibility gene and a follow-up study[98] of the 3q peak demonstrated a significant association of diabetic nephropathy to IL20RB-NPMIP17 gene region on that chromosomal. Other chromosomal regions have also been associated with diabetic nephropathy in different populations. For example, the Family Investigation of Nephropathy and Diabetes (FIND)[99], looking at the predisposition for nephropathy in diabetic sibling pairs concordant and discordant for diabetic nephropathy, found the strongest evidence of linkage to the diabetic nephropathy trait on chromosomes 7q, 10p, 14q and 18q. Subsequently, these have been replicated by genome-wide linkage studies in other populations. For example, the locus on chromosome 18q demonstrated in the FIND study has also been linked to diabetic nephropathy in African–Americans[100] with diabetes, Pima-Indians[101] and African American populations[102]. While the linkage studies have helped in understanding the genetic architecture of DKD and confirmed the genetic predisposition to DKD. However, these studies are non-specific and identify large regions of genome associated with the DN phenotype, rather than identifying the specific locus/variant. Hence, the advent of cheaper genotyping technologies has seen a phasing out of linkage studies and has made GWAS an investigation of choice for identifying susceptibility loci for DKD.

Table 5: Genome wide Association studies using Sib Pair Designs for Type 1 and Type 2 Diabetic Nephropathy

	Authors	Year	Ethnicity	Study Type and Sample Size	Phenotypes	Linkage Peaks
Type 2 Diabetes	Imperator et.al[103]	1998	Pima Indians	Sibling Pairs	Proteinuria and ESRD	7q, 3q
	Vardarlu et.al[101]	2002	Turkish	18 Turkish families(368 subjects)	Proteinuria	18q
	Bowden et.al[100]	2003	African Americans	206 affected sib pairs from 166 African American families (355 affected individuals)	ESRD	3q,7p,18q
	Freedman et.al[44]	2005	Caucasians		ESRD	13q, 9q,4p,1q
	Krowleski et.al[104]	2006	Caucasians	63 extended families with an average of 6.8 diabetic members (range 2–14) and 6.8 nondiabetic members (range 1–18) per family	ACR	5q,7q,22p
	Placha et.al[70]	2006	Caucasians	63 extended families 406 type 2 diabetes	eGFR	2q,10q,18p,3q,7p
	Iyengar et.al[105]	2007	Caucasians, African American and Mexican Americans	1,227 participants from 378 pedigrees	Proteinuria and ESRD and ACR	10p,14q,18q,2q,7q,15q
	Chen G et al[106]	2007	West Africans	321 sib pairs and 36 half-sib pairs, 691 patients	Serum creatinine, Creatinine clearance and eGFR	16q24,7q
	Freedman et.al[107]	2008	Caucasians and African Americans	eGFR, ACR	creatinine ratio (ACR), serum creatinine	2p,7q,13q,3p,10p
	Igo Rp et.al[99]	2011	African-American, American-Indian, European-American and Mexican-American	1,235 nuclear and extended pedigrees (3,972 diabetic participants)	urine albumin: creatinine ratio (ACR)	6p,7p,3p,7q,16q, 22q
Type 1 Diabetes	Moczulski et.al[108]	1998	Caucasians	66 discordant sib pairs for DN	Proteinuria, ESRD	3q
	Osterholm et.al[109]	2007	Finnish	83 discordant sib pairs for DN	Proteinuria	3q
	Rogus et.al[110]	2008	Caucasians	100 discordant sibpairs for DN	Proteinuria and ESRD	19q, 6p, 6q, 3q

Genome-wide Association Studies for Diabetic Nephropathy

The development of advanced DNA sequencing technologies and improvements in computational power to analyse large-scale data has changed the genetic investigation of common complex disorders. GWAS operate on the assumption of "common variant common disease" hypothesis[111], which predicts that common disease-causing alleles - or variants - will be found in all human populations which manifest a given common disease. Since the prevalence of DKD is common (up to 40%, USRDS, 2003) in both T1D and T2D patients, GWAS are considered as an appropriate tool to investigate the common variants associated with DKD. Over the last decade, several GWAS have attempted to identify the key loci associated with T1D nephropathy and T2D nephropathy. We will now review some of these some of these GWAS studies with adequate sample sizes and a positive replication.

GWAS for T1D Nephropathy in Genetics of Kidneys in Diabetes Collection (GoKinD)

This study[112] genotyped approximately 360,000 SNPs in 820 case subjects (284 with proteinuria and 536 with end-stage renal disease) and 885 control subjects with Type 1 diabetes Affymetrix 5.0 500K SNP array and sought replication in 1,304 participants of the Diabetes Control and Complications Trial (DCCT)/Epidemiology of Diabetes Interventions and Complications (EDIC) study[113] - a long-term, prospective investigation of the development of diabetes-associated complications. Case subjects with diabetic nephropathy had either persistent proteinuria, defined by a urinary ACR ≥300 μg/mg in two of the last three measurements taken at least one month apart, or ESRD (dialysis or renal transplant). Controls for DKD had Type 1 diabetes for at least 15 years and normo-albuminuria, (defined by an ACR <20 μg/mg in two of the last three measurements taken at least one month apart without

any antihypertensive treatment). The strongest with nephropathy association was at the FRMD3 (odds ratio [OR] = 1.45, P = 5.0 x 10(-7)) and CARS) locus (OR = 1.36, P = 3.1 x 10(-6)). Associations between both loci and time to onset of diabetic nephropathy were supported in the DCCT/EDIC study (hazard ratio [HR] = 1.33, P = 0.02, and HR = 1.32, P = 0.01 respectively). This study, however, had several limitations. None of the markers tested for association with DKD reached genome-wide significance in the discovery cohort. The discovery cohort was a cross-sectional case-control sample while replication was sought in a prospective survival analysis model looking at the time to onset of severe nephropathy in the DCCT/EDIC study cohort. Due to differences in study design in the discovery and replication cohorts, the authors were unable to report on the joint effect of the two studies. Another limitation of the study is the high proportion of patients with ESRD (60%) in the discovery cohort. This limited the ability of the study to detect variants associated with proteinuria. In addition, since the patients with ESRD had survived on dialysis or transplant for several years, a variant, which can increase mortality or cause a severe ESRD, can go undetected.

GWAS for Type 1 Diabetic Nephropathy in American European Population (Genie Study)

In the Genetics of Nephropathy: an International Effort (GENIE) consortium[97], a meta-analysis of genome-wide association studies (GWAS) of T1D DN comprising ~2.4 million single nucleotide polymorphisms (SNPs) imputed in 6,691 individuals from the UK-ROI, GOKIND US and FinnDiane cohort. The study genotyped 41 top ranked SNPs representing 24 independent loci in and additional sample of 5,873 individuals. Combined meta-analysis revealed association of two SNPs with ESRD: rs7583877 in the AFF3 gene (P = 1.2×10−8) and an intergenic SNP on chromosome

15q26 between the genes RGMA and MCTP2, rs12437854 (P = 2.0×10−9). Functional data reported in the study suggested that AFF3 influences renal tubule fibrosis via the transforming growth factor-beta (TGF-β1) pathway. This AFF3 locus, however, was driven by two cohorts and technically did not replicate (p=0.25 in stage 2 replication), although the direction of effect was consistent across studies. The strongest association with DKD (any albuminuria *vs.* no albuminuria) as a primary phenotype was seen for an intronic SNP in the ERBB4 gene (rs7588550, P = 2.1×10−7). Although both AFF3 and ERBB4 represent new signals in the pathogenesis of DN lack of significant association in replication analysis calls for independent confirmation of this locus in other studies before their implication for DN pathogenesis can be examined.

GWAS for Type 2 Diabetic Nephropathy in African American Population (McDnough 2010)

This study[114] genotyped 832,357 autosomal SNPs in 965 Type 2 diabetic African American patients with end-stage renal disease (ESRD) and in 1029 African Americans without Type 2 diabetes or kidney disease as controls on an Affymetrix 6.0 platform. Patients with T2DM-ESRD were recruited from dialysis facilities at the Wake Forest University. T2DM was diagnosed in African-Americans who reported developing diabetes after the age of 25 years and who did not receive only insulin therapy since diagnosis. Cases had T2DM diagnosed at least 5 years before initiating renal replacement therapy, background or greater diabetic retinopathy, and/or ≥100 mg/dl proteinuria on urinalysis in the absence of other causes of nephropathy (T2DM-ESRD subjects). Unrelated African American controls without a current diagnosis of diabetes or renal disease were recruited from the community and internal medicine clinics (Control subjects). For replication of the top markers

African American T2DM-ESRD cases and non-diabetic, non-nephropathy controls were recruited using the same criteria as the case and control subjects that were used in the GWAS. The strongest association on a combined analysis of discovery and replication cohort was seen with rs6930576 ($P = 7.04\times10-7$) in the SASH1 gene, rs7769051 ($P = 6.45\times10-6$) in the RAS12 gene, rs773506 ($P = 6.45\times10-6$) in the AUH gene, rs2358944 ($P = 3.54\times10-6$) within the RASAP2 gene. The other SNPs located in the LIMK2 gene reached near GWAS significance rs2106294, rs4820043 and rs5749286 ($P = 4.11\times10-6$ and $5.07\times10-6$, $9.79\times10-6$). The MYH9 SNP seen in the initial GWAS, rs5750250 did not replicate and had a combined P value = $1.66\times10-7$. One of the major limitations of the study design was the fact that it looked at the association of diabetic ESRD versus non-diabetic non-ESRD controls. So the associated SNPs could either be associated with T2D nephropathy or all-cause ESRD or T2D itself. The authors tried to address this issue by genotyping additional 1,246 AA T2DM, non-nephropathy cases and 1,216 non-T2DM ESRD cases to differentiate between T2DM-ESRD, T2DM, and all-cause ESRD loci. The authors hypothesized that T2DM-ESRD SNPs should have allele frequency differences when compared with cases with T2DM alone (lacking nephropathy) and observed that there were difference in the allele frequencies between T2DM-ESRD SNPs and T2DM-non-ESRD SNPs for the top 25 SNPs in the combined analysis, suggesting that there were therefore indeed ESRD associated SNPs. However, since the discovery cohort and the replication cohort had compared diabetic ESRD vs. non-diabetic non-ESRD controls it is possible that the study failed to identify DM specific ESRD loci with modest effect sizes. The study also evaluated preciously reported genetic associations (*PVT1, FRMD3, CARS, ACACB, NEDD4L, SERPINB7, CNDP1, CNDP2, ELMO1, SHROOM3, UMOD, GATM-SPATA5L1,*

GCK2, ALMS1, DAB2, SLC34A1, VEGFA, SLC22A2, PRKAG2, STC1, ATXN2, DACH1, SLC7A9) with DN/ESRD in Caucasians and Asian populations. None of these genes were associated with DKD in African-American population after correction for multiple comparisons indicating either a genetic heterogeneity in the pathogenesis of DKD-ESRD in those with African ancestry or a limited power of the study to detect these genetic effects in an admixed population. Finally, none of the markers in the study reached threshold for GWAS significance suggesting a need for further replication and analysis of these markers in larger cohorts of DKD-ESRD.

GWAS for DKD stratified by MYH9 and APOL1 Risk Variants

This study[115] performed genome-wide association analyses using the Affymetrix SNP Array 6.0 in 966 African Americans with T2DN and 1,032 non-diabetic, non-nephropathy (NDNN) controls, with and without adjustment for *MHY9* and *APOL1* nephropathy risk variants on chromosome 22. The study hypothesized that since[115,116] *MHY9* and *APOL1* are strongly associated with nephropathy in African Americans[116]; these variants could mask the effects of other variants with smaller effect size. MYH9/APOL1 locus was first identified by admixture mapping study for ESRD in African-American population. The admixture mapping studies are based on the hypothesis that the since the African ancestry is a risk factor of ESRD, the regions of genome with significantly high proportion of African Ancestry (as compared to the controls) are more likely to harbour the disease causing gene. This study[116] looked at 1372 cases with ESRD and 806 controls with ESRD and found significantly excess African ancestry for ESRD phenotype on 22q12 region harbouring the MYH9/APOL1 gene.

Stratified analyses based on the chromosome 22-nephropathy risk haplotypes demonstrated that *FRMD3* variants were associated with diabetic nephropathy risk in cases without two *MYH9* (or *APOL1*) risk haplotypes. This effect was replicated in an independent sample of 640 African Americans cases of DN and 683 African American Controls without DKD. An important lesson from the study is that if there is a common haplotype with big effect size on a given phenotype, it may mask the effects at other loci with modest effect sizes. This is important information to consider while designing GWAS studies, where most of the variants tend to have smaller effect sizes and their effects can be masked by a single variant with big effect size. Hence, it is important to do genome-wide conditional analysis to see after the DKD GWAS to identify if a single SNP with a big effect size is masking the effects of other SNPS with smaller effects.

Table 6: Genome-wide Association Study for Diabetic Nephropathy

Study	Phenotype	Initial Sample Size	Replication Size	Gene	Region	OR/Beta	Top P-value	Platform
Maeda,2007[117]	ESRD	94 cases, 94 controls	No replication	ELMO1	7p14.2	2.67	8x10-6	NR [~80,000]
Hanson,2007[118]	ESRD	105 cases, 102 controls	No replication	PVT1	8q24.21	2.97	2x10-6	Affymetrix [115,352]
Craig,2009[119]	ESRD	547 Caucasian cases, 549 Caucasian controls	No replication	ZMIZ1 LOC100132891	 	1.47 1.56	8.1 x 10-5 1.6 x 10-5	Illumina [474,050] (pooled)
Pezzolesi,2009[112]	ESRD	820 cases, 885 controls	1,304 individuals	FRMD3 CARS CPVL/CHN2	9q 11p 7p	1.45 1.36 1.39	6.3 × 10-7 6.4 × 10-7 5.0 × 10-6	Affymetrix [359,193]
McDonough, 2010[114]	ESRD	965 African American cases, 1,029 African American controls	709 African American cases, 690 African American controls	RPS12 SASH1 LIMK2 MSRB3-HMGA2	6q23.2 6q24.3 22q12.2 12q14.3	1.28 1.31 1.75 1.33	2x10-6 7x10-7 4x10-6 4x10-6	Affymetrix [832,357]
Freedman BI, 2011[115]	ESRD	952 African American cases, 988 African American controls	640 African American cases, 683 African American controls	FRMD3	9q21.32	1.28	4.82x10−4	Affymetrix [832,357]
N Sandholm 2012[97]	DN and ESRD	1399 cases, 5253 controls	387 cases/3465 controls	AFF3	2q11.2-q12	1.23	2.04x10-9	Omni1-Quad array /Affymetrix 500

Summary of Genetic Epidemiology studies for Diabetic Nephropathy

Advances in sequencing technologies have led to a series of candidate gene studies linkage studies and GWAS in the last two decades. With these efforts, there is indeed a better understanding of genetic susceptibility to DKD; however, the lack reproducibility of genetic DN studies has been a disappointing factor. For example, some of the obvious candidates for DN such as the *ACE, APOe* and *TGFβ2* have not replicated in the GWAS studies so also are the established loci from the sib-pair analysis; while most of the GWAS loci do not fall in the known linkage peaks. There can be several reasons for this non-replication.

The variable definition of DKD in genetic epidemiology studies can explain the overall disappointing reproducibility of genetic DKD studies. There are two important aspects to diabetic kidney disease- diabetic albuminuria and renal function decline. DKD is characterised by intersecting stages of declining GFR and progressive proteinuria, both influencing each other[120] finally leading to ESRD. However, most of the studies looking at the DKD focus only on one aspect of the disease – mostly albuminuria. Some studies have used the most severe forms of DKD (ESRD) to define cases reduces some potential misclassification of phenotype but still this does not overcome the critical inaccuracy of case definition not based on histology. In addition, cases with ESRD are by definition survivors (although nephropathy is also the leading risk factor for premature mortality at population level); and by modelling ESRD these studies might be identifying genes protective for DKD mortality. One way of addressing this issue in GWAS would be modelling all the intermediate phenotypes in DKD such a micro-albuminuria, macroalbumiuria and ESRD and adjusting for eGFR while modelling albuminuria in GWAS studies of DKD.

Chip-based heritability can now be estimated using GWAS data set quantifying the additive genetic variance explained by all SNPs in GWAS datasets[121] and this information can be used to make a priori selection of DKD phenotypes. Estimation of chip-based heritability with GCTA needs large sample sizes-and can be computationally challenging, however, this approach can be used to fine-tune the phenotypic definitions for DKD.

Another problem with trying to identify genetic loci for diabetic nephropathy in Type 2 diabetes is that in most studies it is difficult to obtain information regarding the major risk factors – duration of diabetes and duration of hyperglycaemia. Most people with Type 2 diabetes are diagnosed years after the disease begins: therefore, there is great inaccuracy in diabetes duration prior to the development of diabetic nephropathy. This has detrimental effects on the measurement of long-term glycaemia – if we do not know when the diabetes and glycaemia began, we cannot adequately adjust of the most important determinant of DN in genetic models. These can reduce the power to estimate true genetic effects.

Finally, most of the candidate genes and GWAS until date have been underpowered due to the limited sample sizes. The reductions in the costs for genotyping technologies and international collaborations can overcome this problem and adequately power GWAS meta-analysis can provide adequate power to identify the DKD susceptibility variants.

In spite of the shortcomings and challenges of the GWAS studies in DKD, it is clear that GWAS studies with adequate sample size, precise phenotype definitions, and well-defined analysis plan can identify novel loci associated with DKD and pursuit of these loci can provide new insights into the pathogenesis of DKD. These insights

in turn will help in designing methods for screening prevention and treatment of DKD.

Aims

The broad aim of this thesis is to increase our understanding of the genetic determination of diabetic kidney disease. This will be achieved by the following sub-aims:

I. Estimating the chip-based heritability of various DKD sub-phenotypes

Diabetic kidney disease can be classified into various stages based on the levels of albuminuria excretion and estimated Glomerular filtration rate (eGFR). These sub-groups can represent various stages in pathogenesis of DKD, which in turn can have different genetic and environmental determinants. There are no heritability estimates for genetic determinants of these albuminuria and eGFR based subgroups. The aims of this analysis are

1) to quantify and compare chip based heritability of various DKD subgroups using a GWAS dataset in patients with Type 2 diabetes
2) to conduct a bivariate analysis to estimate shared heritability between blood pressure and DKD, and HbA1c and DKD using a GWAS dataset in patients with Type 2 diabetes

II. Replication of known SNPs for upstream risk factors for DKD (HbA1c and BP) and look at the cumulative effect of these on DKD

HbA1c and BP are known upstream risk factors for DKD. Several GWAS studies have identified loci associated with HbA1c and BP. To what extent diabetes status

attenuates or exaggerates these associations with HbA1c and BP is unknown. The aims of this analysis are

1) to establish the association of known GWAS identified SNPs for HbA1c and BP in patients with Type 2 diabetes
2) to study the cumulative effect of HbA1c and BP associated SNPS on DKD using genetic risk score analysis in patients with Type 2 Diabetes

III. Replication of the known loci for eGFR in general population in patients with T2D and identifying loci for CKD by contributing to the international consortia studying genetics of CKD (CKDGen)

GWAS studies have identified SNPs associated with eGFR in general population. However, there are no independent replications exclusively in patients with Type 2 diabetes. The aim of this study is

1) to replicate the association of eGFR associated loci in general population in patients with T2D
2) to estimate the effect of albuminuria on association of eGFR with known SNPs for eGFR
3) to contribute data to the CKDGen consortium for replication of their top SNPs

IV. Performing a GWAS for known biomarker (sRAGE) for DKD

There is a growing body of evidence suggesting that AGE-RAGE axis is involved in the pathogenesis of diabetic nephropathy. However, very little is known about the

genetic determinants of the circulating form of RAGE (sRAGE). Given its role in DKD, it is important to identify if sRAGE is a genetically fixed trait or if it can be modified by environment and drugs. The aims of this study are

1) to estimate the heritability of sRAGE
2) to do a GWAS to identify Genetic determinants of sRAGE

V. Performing genome-wide meta-analysis to identify novel SNPs associated with DKD

High heritability of DKD and limited GWAS data for DKD suggests that GWAS studies with adequate sample size, precise phenotype definitions, and well-defined analysis plan can identify novel loci associated with DKD. The aims of this study re

1) to do GWAS analysis for various DKD phenotypes for Type1 and Type 2 Diabetes
2) to combine data in meta-analysis of GWAS datasets to identify novel determinants of DKD

Methods

I. Datasets used in the thesis

Three datasets, namely Genetics of Diabetes Audit and Research Tayside `[122,123], EURODIAB[124], and Collaborative AtoRvastatin Diabetes Study (CARDS)[125] were used for statistical analysis in this thesis. This chapter briefly describes the data sources, quality controls procedures, and generic and specific analysis methods for each of the three datasets.

II. DARTS and Go-DARTS

CHI-Master Index and SMRO1 and GRO data

The Health Informatics Centre (HIC) in partnership with the University of Dundee, National Health Service Tayside, and the information services division of national services provides researchers and others with information derived from person-specific datasets. These datasets originate from data held by the University of Dundee and the National Health Service and are anonymised in accordance with the Standard Operating Procedures approved by the Caldicott Guardians. In Scotland, every person registered with a medical practitioner is assigned a Community Health Number (CHI). CHI is a unique 10-digit identification number, which is linked to personal information (e.g. address, postcode, and date of birth) and clinical data recorded during health care activities from primary to tertiary care including a record of medical practitioner registration status, date of birth and date of death. In Tayside this information pertaining to the entire population is held by the Tayside Health Board. Tayside also uses this number thus allowing for the record-linkage of datasets.

The Diabetes Audit and Research in Tayside Scotland (DARTS) research database includes information for all patients diagnosed with diabetes in Tayside (Figure 1). Individuals with diabetes were identified from hospital records[126]. This database was validated against general practice records and confirmed to be robust. Furthermore, the methodology used was shown to be more sensitive for the identification of individuals with diabetes when compared with the use of general practice records alone[126].

Participants with T2D were identified for enrolment through DARTS - a comprehensive and well-validated region-wide clinical information system for diabetes that incorporates multiple clinical data sources[127]. Age- and sex-matched diabetes-free participants were identified in populations within the region of Tayside from general practice records. Taken together these two populations form the Go-DARTS cohort which includes 17,602 participants enrolled between December 1998 and May 2009 and in which there are approximately equal numbers of participants with T2D (N=9,829) and without T2D (N=7,773).

Relevant clinical data available for the Tayside region for all Go-DARTS participants were drawn from electronic records of hospital admissions (Scottish Morbidity Register, SMR01), deaths (General Registry Office, GRO), biochemical tests, and dispensed drug prescriptions. Data were available from 1980 until present for the SMR hospital admissions data; from 1998 for deaths from the GRO; from 1993 until present for prescriptions; and from 1980 until present for biochemical tests.

The Tayside Committee for Medical Research Ethics approved the Go-DARTS study and written informed consent was obtained from each participant. A single sample of blood was collected for DNA extraction and genotyping and the

participant was assigned a unique anonymised system identifier. Baseline characteristics were recorded at the time of recruitment for all participants[127].

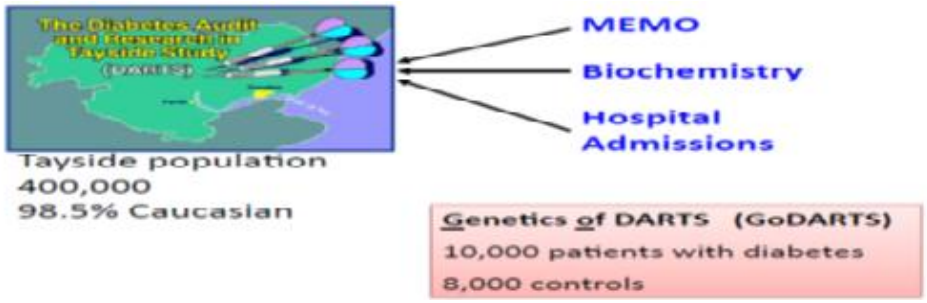


Figure 1: DARTS and Go-DARTS study

Laboratory Data for eGFR and Albuminuria (ACR and MA) and Renal Phenotypes

Data for eGFR (serum creatinine) and ACR (albumin creatinine ratio) were extracted from the Laboratory Biochemistry file and the SCI-DC Biochemistry file available in the Go-DARTS dataset.

eGFR was estimated using the MDRD equation-Modification of Diet in Renal Disease Study Group[128]. This equation estimates eGFR using four variables: serum creatinine mg/mmol, age in years, ethnicity, and gender.

MDRD Equation is given as

((GFR = 175 x $SerumCr^{1.154}$ * $age^{0.203}$ * 1.212 (black) * 0.742 (if female))

These MDRD equations are to be used only if the laboratory has *not* calibrated its serum creatinine measurements to isotope dilution mass spectrometry (IDMS). When IDMS-calibrated serum creatinine is used (which is about 6% lower), the above equations should be multiplied by 175/186, or by 0.94086. There can be variations in the creatinine measures over the period of time due to differences in the assays used by lab (IDMS-traceable Jaffe and enzymatic creatinine assays),

however, the SUMMIT and other studies looking at creatinine measures in this thesis, involved analysis of most recent cross-sectional creatinine levels. Hence, the variations in the assays is unlikely to affect the analysis.

Other equation to estimate eGFR from serum creatinine includes the CKD-EPI[129] (Chronic Kidney Disease Epidemiology Collaboration). This was developed in an effort to create a formula more accurate than the MDRD formula, especially when actual GFR is greater than 60 mL/min per 1.73 m2.

CKD-EPI Equation is given as

$GFR = 141 \times \min(Scr/\kappa,1)^{\alpha} \times \max(Scr/\kappa,1)^{-1.209} \times 0.993^{Age} \times 1.018$ [if female] X 1.159 [if black]

Where Scr is serum creatinine (mg/dL), κ is 0.7 for females and 0.9 for males, α is –0.329 for females and –0.411 for males, min indicates the minimum of Scr/κ or 1, and max indicates the maximum of Scr/κ or 1.

The CKD-EPI equation performs better than the MDRD equation, especially at higher GFR, with less bias and greater accuracy[130]. However, we chose to use the MDRD equation to maintain uniformity in eGFR estimation across various cohorts in the multi-centre GWAS meta-analysis as some of the centres only had MDRD eGFR Also the CKD-EPI performs better only in those with higher eGFR[131]. Since most of our study participants were older patients with Type 2 diabetes diagnosis over a long duration, they are expected to have a lower eGFR at baseline and MDRD is expected to perform well at these lower eGFRs.

Two measures of albuminuria were available for classification of patients in those who had albuminuria and those who did not: 1) Macro-albumin (MA) test - is a test conducted routinely in the diabetes clinic at Ninewells hospital, which measures albumin on spot urine. 2) Albumin creatinine ratio (ACR) - a test conducted only in

a subset of patients attending the diabetes clinic. These ratios correct for variations in urinary concentration due to hydration and provide a more convenient method of assessing protein and albumin excretion than that involved with timed urine collections (measurement of protein excretion in a 24-hour collection). Based on the review of evidence accumulated over three decades, the KDOQI Work Group[6] recommends the use of "spot" urine measurements that compares the concentration of protein to the concentration of creatinine.

Since ACR data was not available on all patients in 2010 we wanted to evaluate the Macro-albumin (MA) test results to see if these can be used to define phenotypes for the GWAS analysis. One way to do this was to compare the albuminuria status of patients at their first available Macro-albumin (MA) test result with their last available Macro-albumin (MA) test result. This was to see if the results remained consistent over time. Table 7 below compares the albuminuria status at baseline versus the albuminuria status at the last follow up using Macro-albumin (MA) test result.

Table 7: Comparison of albuminuria status at baseline versus the albuminuria status at the last follow up using Macro-albumin (MA) test result

	Albuminuria status at last MA reading				
Albuminuria status at First MA reading	**Controls**	**High-micro**	**Macro**	**Micro**	**Total**
Controls(<20 mg/L)	7013(90%)	50 (29%)	39 (28%)	878 (55%)	7980
Highmicro (≥100 < 200mg/L)	53(0.06%)	48 (27%)	20 (14%)	63 (3%)	184
Macro (≥200mg/L)	18(0.02%)	12 (6%)	56 (40%)	28 (1%)	114
Micro (≥20 < 100mg/L)	677 (8%)	62 (36%)	23 (16%)	625 (39%)	1387
Total	7761	172	138	1594	9665

Table 8: Comparison of Albuminuria status at last follow-up using Albumin Creatinine ratio and Macro-albumin test

ACR-status	MA status				
	Controls	**Highmicro**	**Macro**	**Micro**	**Total**
Controls(< 2.5)	4080 (84.5%)	5 (5%)	4 (6%)	300 (32%)	4389
Micro (≥20 < 100mg/L)	558 (11.5%)	30 (32%)	5 (8%)	409 (44%)	1002
Highmicro (≥20 < 100mg/L)	84 (1.5%)	29 (31%)	13 (20%)	108 (11%)	234
Macro (≥200mg/L)	87 (1.5%)	27(29%)	40 (64%)	102 (11%)	256
Total	4809	91	62	919	5881

The above Table 7 shows that about 9% of patients who had albuminuria at baseline were classified as controls at their last follow-status with Macro-albumin (MA) test result. There is also a misclassification in terms of albuminuria subgroups. For example, 24% of those who have macro-albuminuria at baseline were classified as having micro-albuminuria at the last follow-up. In Table 8, we compare the albuminuria status in those who had both ACR and MA test results. Of the total 4,809 patients with diabetes classified as having normo-albuminuria by MA test, 11% had micro-albuminuria, 1% had high micro-albuminuria and 1% had macro-albuminuria based on their status on ACR test. Given the fluctuations in the Macro-albumin (MA) test result and poor agreement with ACR data, a decision was taken to request ACR for all the patients in Ninewells Diabetes clinic starting year 2011. The final phenotype for GWAS analysis was defined exclusively by the ACR test results in patients with Type 2 Diabetes.

Genotype Data

The quality of the genomic DNA was validated using the Sequenom iPLEX assay designed to genotype four gender SNPs and 26 SNPs present on the Illumina Beadchips. DNA concentrations were quantified using a PicoGreen assay (Invitrogen) and an aliquot assayed by agarose gel electrophoresis. A DNA sample was considered to pass quality control if the DNA concentration was $\geq$ 50 ng/µl, the DNA was not degraded, the gender assignment from the iPLEX assay matched that provided in the patient data manifest, and genotypes were obtained for at least two thirds of the SNPs on the iPLEX.

Genotype data quality control of the discovery samples was similar to other Wellcome Trust Case Control Consortium 2 (WTCCC2) studies published elsewhere[132].

For all individuals, we explicitly modelled the data as a mixture of 'normal' and 'outlier' individuals for each of ancestry, missing data, heterozygosity, and sex assignment. We fitted each model in a Bayesian framework and excluded individuals whose posterior probability of belonging to the outlier class was above 0.5145. This approach replaces the traditional concept of fixed exclusion thresholds for parameters such as call rate, heterozygosity and ancestry.

To assess relatedness among study individuals, we compared each individual with the 100 individuals they were most closely related to (on the basis of genome-wide levels of allele sharing) and used a Hidden Markov Model (HMM) to decide, at each position in their genome, whether the two individuals shared 0, 1 or 2 chromosomes Identical By Descent (IBD). This allowed a more refined assessment of the relatedness between individuals than genome-wide sharing statistics (for example, parent-child relationships can be distinguished from those of siblings). Individuals were removed from the study iteratively to ensure there was no pair of individuals with IBD≥ 5%. Within each pair of putatively related individuals, the individual with more missing genotypes was removed.

Four thousand diabetic cases were genotyped on the Affymetrix 6.0 SNP genotyping array that includes 1,000,000 SNPs. These individuals were specifically chosen for genotyping as they had all gone on to receive statins after recruitment to Go-DARTS.

Four thousand diabetic cases were genotyped on the Illumina Omni-express array, which consists of ~ 700K SNPs, selected from Hap Map 1-3 for SNPs with a MAF

greater than 5%. The array was designed by selecting tag SNPs to serve as a proxy for a number of others SNPs across the genome. This approach allows for the broadest selection of maximally informative markers, resulting in genome-wide coverage of both common and rare variants.

III. CARDS Clinical Trial

The Collaborative Atorvastatin Diabetes Study (CARDS) is a multicentre; placebo-controlled, double-blind study that enrolled 2,838 white men and women aged between 40 and 75 years of age and randomised them to receive 10 mg/day of atorvastatin or placebo[125]. Patient selection was dependent on a positive diagnosis of Type 2 diabetes in addition to at least one other risk factor for coronary heart disease. The aim of the Collaborative Atorvastatin Diabetes Study (CARDS) was to assess the effectiveness of 10mg of atorvastatin daily versus placebo in the primary prevention of cardiovascular disease in patients with Type 2 diabetes.

Investigators in 132 clinical centres around the UK and Ireland identified potentially eligible individuals by reviewing computerised registers of patients and by opportunistic assessment of people attending diabetes clinics. Men and women aged 40–75 years with Type 2 diabetes mellitus (defined with 1985 WHO criteria) diagnosed at least 6 months before study entry were considered for inclusion, provided they had at least one or more of the following: a history of hypertension, (defined as receiving antihypertensive treatment or having systolic blood pressure of 140 mm Hg or greater or, diastolic blood pressure of 90 mm Hg or greater on at least two successive occasions); retinopathy(any retinopathy; maculopathy or previous photocoagulation); micro-albuminuria or macro-albuminuria (defined as a positive Micral or other strip test; an albumin creatinine ratio of 2•5 mg/mmol or greater; or

an albumin excretion rate on timed collection of 20 μg/min or more, all on at least two successive occasions); or currently smoking (no minimum number of cigarettes per day was required). All patients reporting current smoking were counselled to quit. Patients were ineligible if they had any history of myocardial infarction, angina, coronary vascular surgery, cerebrovascular accident, or severe peripheral vascular disease (defined as warranting surgery). The clinical trial checked eligibility against the patient's clinical notes and their own recall and assessed lipid eligibility criteria by blood testing at one screening and four pre-treatment visits over a 10-week period. Patients were asked to attend these visits after a 12h fast. Mean serum LDL-cholesterol concentration during baseline visits had to be 4•14 mmol/L or lower and serum triglycerides 6•78 mmol/L or less. Patients were excluded if they had a plasma creatinine concentration >150 μmol/L, glycated haemoglobin (HbA1c) of >12% or, if during the baseline phase they had <80% compliance with placebo. Patients were randomised between November 1997, and June 2001.

CARDS Phenotype Data

For the GWAS analyses, patients who were randomised to atorvastatin were included. Covariate information was also extracted e.g. age, sex, BMI sRAGE and esRAGE levels at baseline.

CARDS Genotype Data

This work was not done during the thesis and is a supplementary information relevant to genotyping of CARDS dataset. No genotyping was done in this thesis project(It was performed by Perlegen). For CARDS, DNA was extracted from whole blood EDTA samples. DNA was isolated from 10ml of frozen blood using the Gentra Puregene DNA Isolation Kit from Qiagen (Cat no. 158389), USA.

Briefly, RBC was lysed with an anionic detergent in the presence of a DNA stabilizer which limits the activity of intracellular DNases. WBC was collected by centrifugation at 2000g for 2 min. RNA was removed by treatment with RNase A. Protein was removed by salt precipitation (centrifugation at 2000g for 5 min). Genomic DNA was recovered by precipitation with isopropanol and centrifugation at 2000g for 5 min, the DNA pellet was washed in 70% ethanol, air dried, and dissolved in hydration solution (1 mM EDTA, 10 mM Tris•CI pH 7.5). Purified DNA was stored at -20°C. DNA aliquots were genotyped at Perlegen Sciences using a proprietary SNP set comprising 599,164 SNPs. 243 SNPs that had discrepant map positions between HapMap and Perlegen were dropped; 517,746 SNPs at which the call rate was >80% and at which the p-value for a test of deviation from Hardy-Weinberg equilibrium (HWE) was at least 10-5 were retained in the analysis. SNP annotation was based on build 36 of the Human Genome Sequence. All SNPs were used in the analysis regardless of allele frequency but the allele frequency was considered when evaluating putative associations. Allele frequency was below 1% at 6% of SNPs. Samples were selected from those people who had been allocated atorvastatin 10mg daily, had given consent for genotyping and had a sample SNP call rate >80%. After applying the exclusions of HWE, we estimated relatedness with PLINK and those individuals with Pi_HAT > 0.25 (excluding first and second-degree relatives) were removed (n=0). Only LDL-c values from time points at which the person was compliant with atorvastatin (based on pill count >80/%) were used.

IV. EURODIAB Study

A cross sectional survey of 3,250 people with IDDM in 29 centres in Europe was funded by the EC in 1989-91 and a follow-up study of this cohort was funded by biomed program in 1996 (The EURODIAB Prospective Complications Study or EURODIAB PCS)[133]. The baseline examination included 3,250 patients (1,668 men and 1,582 women; mean [±SD] age, 32.7±10.2 years; mean duration of diabetes, 14.7±9.3 years). These patients were randomly selected in a stratified manner from 31 diabetes clinics across Europe[134].

EURODIAB Phenotype data

The EURODIAB phenotype data for SUMMIT studies was evaluated based on presence or absence of dialysis, serum creatinine, and timed 24-hour urine collection. Cases for the study comprise patients with Type 1 diabetes on dialysis or with macro-albuminuria (any micro-albuminuria ≥ 300 mg / 24 hours on a timed urine collection) or elevated serum creatinine (>200 µmol/lit) consistent with ESRD. For SUMMIT studies, we added cases of persistent micro-albuminuria (i.e. any micro-albuminuria at both baseline and follow up in EURODIAB PCS) and high micro-albuminuria (any micro-albuminuria ≥ 100 mg / 24 hours on a timed urine collection). As described above we captured cases from several sources (EURODIAB at baseline, EURODIAB at follow up, additional cases from these centres not in the original cohort study and renal failure cases from several new non-EURODIAB centres). Controls were only recruited from the original EURODAIB IDDM Complications Study cohort. These Type 1 DM patients had at least 15 years of Type 1 Diabetes remained normo-albuminuric for that period. In addition to local

MICRAL strip testing, they have normo-albuminuria confirmed by the central EURODIAB on two overnight collections at follow up and on one collection at baseline.

EURODIAB Genotype data

1,231 T1D diabetic cases were genotyped on the Illumina Omni-express array, which consists of ~700K SNPs, selected from Hap Map 1-3 for SNPs with a MAF greater than 5%. The array was designed by selecting tag SNPs to serve as a proxy for a number of others SNPs across the genome. This approach allows for the broadest selection of maximally informative markers, resulting in genome-wide coverage of both common and rare variants.

EURODIAB Genotype data quality control

Since EURODIAB is an old study with DNA preserved for more than 10 years, we performed sensitivity analysis on the genotyped data by a) comparing the GWAS data with previously genotyped EURODIAB data for candidate genes and b) testing the GWAS data in well-characterised variants and trait.

The concordance rate between randomly selected GWAS SNPs and the in-house generated GWAS data was about 96%. Furthermore, association of APOe2 variant was examined with baseline lipids and FTO variant with BMI. The direction of effect for this very well characterised genotype-phenotype correlation was consistent with previously reported effects in the literature.

Twenty individuals showed gender mismatch gender reported from the Sequenome run as compared to reported gender. These individuals were removed from subsequent analysis.

Table 9: Concordance rate between the GWAS sNPS common to the prior candidate SNP EURODIAB data

Sr number	SNP	Total genotype count	Genotype matched in GWAS and old EURODIAB data	Genotype miss-match in GWAS and old EURODIAB data	% agreement
1	rs10318GWASA	679	648	31	95.43
2	rs17168032GWASC	678	658	20	97.05
3	rs1869261GWASG	665	639	26	96.09
4	rs1881538GWASA	679	651	28	95.88
5	rs1982436GWASG	706	682	24	96.6
6	rs2270812GWASA	693	670	23	96.68
7	rs3194515GWASA	697	667	30	95.7
8	rs3807337GWASG	692	659	33	95.23
9	rs3812934GWASG	694	666	28	95.97
10	rs6467557GWASA	702	677	25	96.44
11	rs6555055GWASC	701	683	18	97.43
Overall agreement between EURODIAB GWAS data and old EURODIAB data					96%

Table 10: Testing the GWAS data in well characterised variants and traits

Association of APOe2 variant with baseline LDLc levels						
CHR	SNP	modelled allele	TEST	NMISS	BETA	P
19	rs7412(APOe2)	A	ADD	174	-0.5199	0.005995
Association of FTO variant with BMI						
CHR	SNP	Modelled Allele	NMISS	BETA	SE	P
16	RS8050136 (FTO)	A	872	0.4178	0.1744	0.01681

V. Statistical Analysis

The software programs for statistical analysis and genotype imputation were SAS 9.2, R IMPUTE2[135], SHAPE-IT[136], GTOOL[137] (J. Marchini) and SNPTEST[138]. The analyses were performed on the Linux platform. SHAPE-IT was used to phase the data before imputations, Impute 2 was used for imputations, GTOOL was used for manipulation of imputed data, and SNPTEST was used to perform association analysis on the imputed datasets.

VI. Genotype Imputation

Imputation takes place in two stages: the first is the estimation of haplotypes from the study population and the second is the imputation of genotypes by comparing study haplotypes to reference panel haplotypes.

VII. Haplotype Inference

There are many methods for haplotype inference in unrelated populations[139] that estimate haplotypes with varying accuracy. The segmented haplotype estimation and imputation tool (SHAPE-IT, http://www.shapeit.fr/) is the recommended method to estimate haplotypes for downstream imputation with IMPUTEv2. The method is highly accurate and computational light when compared with other available methods[136] and is particularly suited to populations that contain high linkage disequilibrium in their genomes such as Caucasians[140]. The inference of haplotypes is computed in a similar way to Phase v2 where all possible haplotypes are estimated from the available genotype set with an associated probability.

VIII. Imputation with IMPUTEv2

"IMPUTEv2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) uses the estimated haplotypes from SHAPE-IT to impute genotypes from the haplotype set. Imputev2 compares the study population haplotypes with up to two reference panels for the imputation of missing genotypes in the study population. Alleles are imputed into the study population by running a forward-backward algorithm to impute missing alleles with a certain probability. The two sets of haplotypes are compared to each other and missing alleles are imputed into the study panel from the reference panel with certain probability. Certain SNPs will be found in haplotypes together therefore, if one SNP was not present in the study panel, but the haplotypes matched, an allele for that SNP was imputed with a certain probability given the alleles that are present in the reference panel and the alleles that are given in the study panel. So if there is good coverage of a haplotype on a particular chip and the haplotype is present in the reference panel the alleles will be imputed accurately. If haplotypes are sparsely covered or SNPs are not in linkage disequilibrium with any other SNPs to form haplotypes then the alleles may be imputed with low confidence in their accuracy or may be missing all together. Given that, we assume both sets of haplotypes are sampled from populations in Hardy-Weinberg equilibrium the allelic probabilities can be converted to genotypic probabilities".

IX. Cluster Computing

The manipulation of these large genotype files and the imputation procedures are computationally intensive. Parallel computing on a high performance cluster (HPC)

of computers was used to analyse and impute the genetic data. The HPC is managed by a Sun Grid Engine, which is responsible for accepting jobs from users, scheduling and distributing jobs to cores within the HPC. SHAPE-IT supports threading of multiple cores so that they can be used as one processor to perform the pre-phasing steps chromosome by chromosome. Specific Perl programs were written to run the pre-phasing steps on the HPC using the multi-threading capabilities of SHAPE-IT. Each chromosome was imputed by dividing the imputation intervals across 5Mb chunks of each chromosome. Each chunk was distributed to a separate core and specific bash (Shell script) program were written to run each of the imputation steps. The resulting files were combined into single chromosomes using bash scripting. Imputed files were also analysed chromosome by chromosome on the HPC. The analyses programs used for directly and imputed genotypes are described below.

X. Genetic Associations with SNPTEST

Genotypes of directly typed SNPs are assigned using a threshold method at a probability of 0.9. SNPs can be modelled in logistic and linear regressions on a log-additive scale. Similar to non-genetic covariates the effect of the SNP is modelled by steps from one genotype to another. SNPs are coded 0 for homozygotes of the non-effect allele, 1 for one copy of the effect allele in a heterozygote and 2 for a homozygote of the effect allele so the effect of each step translates to the effect of an additional copy of the effect allele on the outcome. Imputed SNPs are more complex to model as the three genotypes are estimated with some probability and the imputed data are assumed to be 'missing' in that they are not observed. Models of association for imputed genotypes need to take into account genotype uncertainty.

SNPTESTv2

(https://mathgen.stats.ox.ac.uk/genetics_software/snptest/old/snptest.html) considers genotype uncertainty by using a missing data likelihood score test. Since SNPs may be imputed with varying certainty, a threshold method may lead to a high proportion of missing data. It is possible to sum the probabilities of each genotype across individuals for a SNP and use all the available data in the association analysis. For a particular SNP data may be missing in some individuals and not in others so the data can be partitioned into observed and missing data. A score test can be used to estimate the likelihood of the observed data given the full set of data and taking into account the missing data.

XI. Principal Component Analysis using EIGENSOFT

The EIGENSOFT package combines functionality from population genetics methods[141] and EIGENSTRAT is a method[142] for correction of population stratification. We used eigenvectors (principal components) obtained from the Principal Component analysis implemented in EIGENSOFT to adjust for population stratification in our datasets.

"Population stratification is the presence of a systematic difference in allele frequencies between subpopulations in a population possibly due to different ancestry, especially in the context of association studies". Population stratification can be a problem for association studies, such as case-control studies, where the association found could be due to the underlying structure of the population and not a disease-associated locus. The two most widely used approaches to this problem include genomic control, a nonparametric method for controlling the inflation of test statistics[143] and structured association methods[144] which use cumulative genetic

information to estimate and control for population structure. Currently, the most widely used structured association method is Eigenstrat, (PCA analysis) developed by Alkes Price.

"PCA is a statistical method for exploring and of datasets with a large number of measurements (which can be thought of as dimensions) by reducing the dimensions to the few principal components (PCs) that explain the main patterns. Thus, the first PC is the mathematical combination of measurements that accounts for the largest amount of variability in the data".

"PCA has a population genetics interpretation and can be used to identify differences in ancestry among populations and samples. In particular, by assessing whether the proportion of the variance explained by the first PC is sufficiently large, it is possible to obtain a formal P value for the presence of population substructure and to identify the number of PCs that are statistically significant. These can then be used to adjust for population stratification in the datasets. However, in practise since the top three PC explain generally explain most of the variation in data, these are using in adjusting for population stratification in regression models".

XII. Example of Principal Component Analysis using the EURODIAB Dataset

Analysis was done by EIGENSOFT SMARTPCA program using, 2,90,517 uncorrelated markers. For this process, we used ten Outlier removal iterations. First three Principal components explained most of the variation in the datasets and were used as covariates to adjust for population structure in the genome-wide association tests. Figure 2 shows the plot of 1st and 2nd principal components on *x* and *y*-axis respectively, the colour coding indicates self-reported ethnicity.

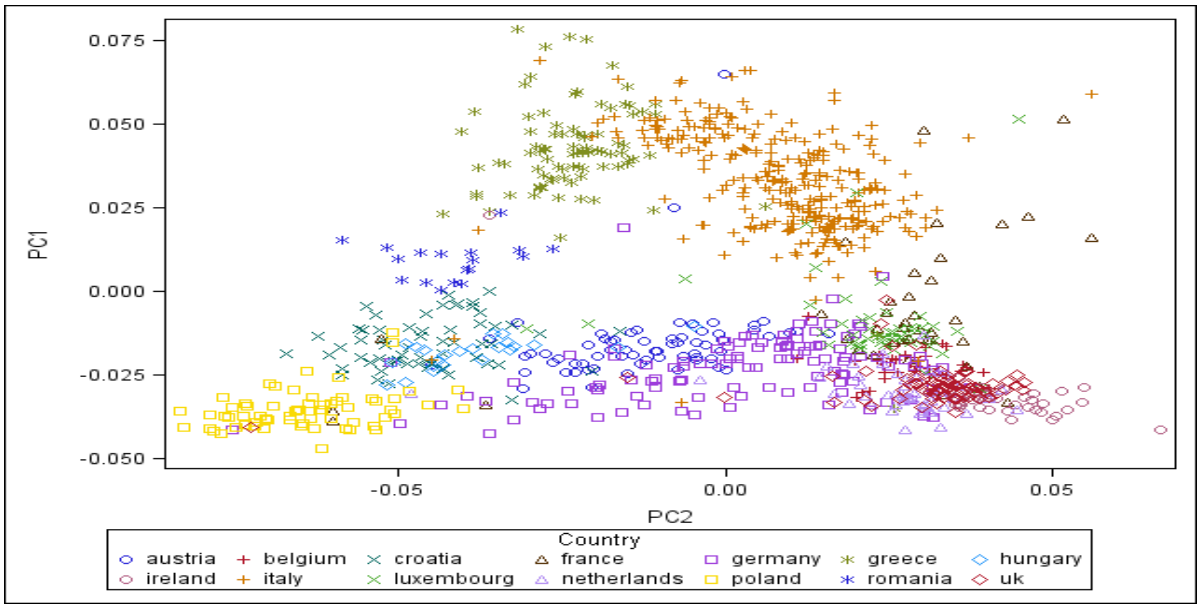


Figure 2: Principal Component Analysis in EURODIAB Dataset

It is evident that people from same ethnicity cluster together in the graph, showing the power of the PCA to cluster individuals using their genotype information.

XIII. Meta-analysis

In a meta-analysis, summary statistics such as effect estimates and the variance of the effect estimates from individual studies are combined. This process allows us to estimate the effects of SNPs across any number of studies without the need to access individual level data. Summary statistics can be combined using a fixed effects model or random effects models[145]. Random effect models were used if there was significant heterogeneity across the estimates.

XIV. Fixed Effects Meta-analysis

The fixed effects model assumes that the effect sizes are the same or fixed across studies. Effects are commonly combined using inverse-variance weighted effect size estimate and the weighted sum of z scores.

$$\beta meta = \frac{\sum Wi\beta i}{\sum Wi}$$

Where Wi =1/(SE)^2 and the $SE_{meta} = \sqrt{\sum Wi}^{-1}$

$Z_{FE} = \frac{Bmeta}{SEmeta}$

A p value can be calculated from the z score assuming a two-sided test:

$P_{FE} = 2\Phi(-|Z_{FE}|)$

Where Φ is the cumulative density function of the standard normal distribution.

XV. Random Effects Meta-analysis

The random effects model assumes that the effect sizes are sampled from a probability distribution that has a variance τ^2 that can be estimated using various approaches such as the method of maximum likelihood and restricted maximum likelihood (Cooper HM The handbook of research synthesis and meta-analysis. New York: Russell Sage Foundation; 2009). The estimated between study variance τ^2 hat is included in the calculation of the effect size estimate where

$$\beta = \frac{\sum(Wi + \hat{\tau}^2)^{-1}\beta\, i}{\sum(Wi + \hat{\tau}^2)^{-1}}$$

$$SE(\beta) = \sqrt{\sum(Wi + \hat{\tau}^2)^{-1}}^{-1}$$

In addition, the z score can be calculated as follows:

$$\text{ZRE} = \frac{\beta}{SE(\beta)}$$

The p value can then be calculated as $P_{RE}= 2\Phi(-|Z_{RE}|)$

The current RE model assumes that there is heterogeneity under the null hypothesis however there should be no heterogeneity under the null hypothesis as all effect sizes are equal to zero. This assumption leads to overly conservative p values.

XVI. Specific Methods for Heritability Estimation using GCTA

GCTA version 1.11 was used to calculate the pair-wise genetic relationship between individuals and create the genetic relationship matrix[146]. Principal components analysis was then applied to all the SNPs to calculate the first 10 eigenvectors, which were included as covariates in all the heritability estimation analyses to control for potential population structure. Univar ate heritability estimation of each definition of diabetic kidney disease phenotype was then performed by restricted maximum likelihood method in GCTA with gender and age at starting at baseline included as covariates.

Statistical significance was determined using the likelihood-ratio test of specific hypothesis. We report the asymptotic 95% confidence interval by 1.96 times the

standard error. As the standard errors of the parameter estimates were derived from first-order Taylor series expansions about the likelihood in GCTA, they may be prejudiced for small study sample size[147], which at borderline levels of significance explains the inconsistency between p-value and 95% CI reported.

In addition, a bivariate analysis was utilised to jointly estimate the heritability of baseline HbA1c, BP and the heritability of DKD. The most informative parameter estimated from such bivariate analysis was the genetic correlation r_g. It represents the proportion of variance shared between baseline HbA1c BP and diabetic kidney disease that was contributed by common genetic determinants.

XVIII. Specific Methods for sRAGE GWAS

CARDS Dataset

The design of the CARDS trial has been reported previously. The trial was conducted in 132 clinical centers in the United Kingdom and the Republic of Ireland. In it, 2,838 patients with Type 2 diabetes without previous CVD were randomised to receive either atorvastatin 10mg daily or placebo. Patients were ineligible if they had any history of myocardial infarction, angina, coronary vascular surgery, cerebrovascular accident, or severe peripheral vascular disease. We checked eligibility against the patient's clinical notes. The median follow up duration in the trial was 3.9 years. The primary endpoint was major cardiovascular disease events comprising myocardial infarction including silent infarction, unstable angina, acute coronary heart disease death, resuscitated cardiac arrest, coronary revascularisation procedures, or stroke.

Using pre-randomisation samples serum sRAGE was measured using the R&D Systems Quantikine Immunoassay that is specific for the extracellular domain of human RAGE (Quantikine; R&D systems, Minneapolis, MN USA). The assay coefficient of variation was 4.4%. esRAGE was measured using the B-Bridge ELISA (B-Bridge International Cupertino, Ca USA). This assay specifically measures the esRAGE/RAGEv1 protein only due to the use of an antibody directed against the unique C-terminus sequence of RAGEv1, and does not cross-react with other potential forms of sRAGE. The coefficient of variation was 6%. Sufficient sample was available for total sRAGE and esRAGE measurement 715 participants. The study utilised data generated by genotyping 2,362 individuals in the CARDS trial. sRAGE level was measured in 587 patients with Type 2 Diabetes and eSRAGE was measured in 645 patients with diabetes .

The CARDS genotype data were combined with phased haplotypes from HapMap phase II CEU r22to compute posterior probability distribution of genotype at all HapMap loci using the IMPUTE program. The genotypes thus generated were used for association testing with sRAGE levels using linear regression implemented in SNPTEST. The association tests were adjusted for age, gender, BMI, population stratification (eigenvectors generated from Principle Component Analysis). In order to normalize the distribution of esRAGE and sRAGE levels the sRAGE and eSRAGE levels were log-transformed.

Go-DARTS Dataset

The Go-DARTS cohort was ascertained from the Diabetes Audit and Research Tayside Study (DARTS). Validated prescribing data, biochemistry data, as well as

clinical phenotypes back to 1992 can be retrieved from central databases for all the DARTS patients. Prospective longitudinal data were also collected on these patients. Since October 1997, all patients with diabetes were invited to give written informed consent to DNA as part of the Wellcome Trust United Kingdom Type 2 Diabetes case control collection. As of June 2009, 8,000 cases and 7,000 controls of European ancestry have participated in this Genetics of DARTS (Go-DARTS) study. sRAGE was measured using the human Quantikine ELISAs (R&D Systems). 50ul of serum was required the measuring range was 0-5000 pg/mL and the Intra-assay and inter-assay CV were 3.89% and 8.5% for sRAGE. In the Go-DARTS dataset sRAGE was measured in 1,193 individuals of whom 455 were genotyped on Affymetrix platform and 463 on the Illumina platform. Three hundred and forty-eight individuals genotyped on the Affy platform passed quality control while 440 individuals genotyped on the illumine platform passed quality control. The Go-DARTS genotype data were combined with phased haplotypes from HapMap phase II CEU to compute posterior probability distribution of genotype at all HapMap loci using the IMPUTE programme. The genotypes thus generated were used for association testing with sRAGE levels using linear regression implemented in SNPTEST. The association tests were adjusted for age, gender, BMI, and population stratification (eigenvectors generated from Principle Component Analysis).

Conditional Analysis of RAGE Gene in CARDS Dataset

We used all the markers in the *AGER* gene region which passed the threshold for significance on the CARDS-GoDARTS GWAS meta-analysis (P-value <10-E-8(n=22)). These were then pruned for LD and only those with LD <20% were

carried into next stage. LD based pruning was done with --indep command. The parameters for --indep are: window size in SNPs equal to 50, the number of SNPs to shift the window at each step equal to 5 and the VIF threshold of 2. The VIF is 1/(1-R^2) where R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously. That is, this considers the correlations between SNPs but also between linear combinations of SNPs. "A VIF of 10 is often taken to represent near collinearity problems in standard multiple regression analyses (i.e. implies R^2 of 0.9). A VIF of 1 would imply that the SNP is completely independent of all other SNPs. The LD pruned markers were then used in a linear regression analysis with sRAGE levels as dependent variables and age, sex, BMI, and population eigenvectors as covariates" (source Plink Manual).

Chip-based Heritability Estimation and Genome-wide Joint Conditional Analysis using Genome-Complex Trait Analysis (GCTA)

Heritability of sRAGE was estimated with GCTA using combined panel of Affymetrix and Illumina datasets in the Go-DARTS study. Summary level statistics from CARDS- Go-DARTS meta-analysis were used to perform a Genome-wide conditional analysis. There may be multiple causal variants in a Gene and the total variation that could be explained at a locus may be undervalued if only the most significant SNP in the region is selected. To identify independent SNPs we ideally can perform a conditional analysis, starting with the top associated SNP, across the whole genome followed by a stepwise procedure of selecting additional SNPs, one by one, according to their conditional *P* values. Such a strategy would allow the discovery of more than two associated SNPs at a locus. To identify independent

SNPs across the genome-wide data we used an approximate conditional and joint analysis approach implemented in GCTA software[148]. We used summary level statistics from CARDS-Go-DARTS meta-analysis and LD corrections between SNPs estimated from CARDS GWAS data. SNPs on different chromosomes or more than 10Mb distant are assumed to be in linkage equilibrium. The model selection process in GCTA starts with the most significant SNP in the single-SNP meta-analysis across the whole genome with *P* value below a cutoff *P* value, such as 5×10^{-8}. In the next step, it calculates the *P* values of all the remaining SNPs conditional on the top SNP that have already been selected in the model. To avoid problems due to co-linearity, if the squared multiple correlations between a SNP to be tested and the selected SNP(s) is larger than a cut-off value, such as 0.9, the conditional *P* value for that SNP will be set to 1. Select the SNPs with minimum conditional *P* value that is lower than the cut-off *P* value. Fit all the selected SNPs jointly in a model and drop the SNPs with the *P* value that is greater than the cut-off *P* value. This process is repeated until no SNPs can be added or removed from the model.

Power Calculation

Power calculations for the study was performed using the software GWAPower[149] GWAPower calculates statistical power using the following input parameters: (1) The heritability (a number in the range 0~1); (2) Total sample size; (3) The number of SNPs in the GWA study; (4) The type 1 error rate: The default value is calculated using a Bonferroni correction. With an approximate discovery sample size of 1,100, our study had 60% power to detect a SNP which explains 3% variation in the levels

of sRAGE and more than 95% power to detect a SNP which explains 5% variation in the levels of sRAGE.

Meta-analysis

Meta-analysis of the GWAS data was performed using Metal[150]. This program selects a reference allele for each marker and calculates a z-statistic characterizing the evidence for association. The z-statistic summarizes the magnitude and the direction of effect relative to the reference allele and all studies are aligned to the same reference allele. An overall z-statistic and p-value are then calculated from a weighted sum of the individual statistics. Weights are proportional to the square root of the number of individuals examined in each sample and selected such that the squared weights sum to 1.0.

XIX. Specific Methods for CKDGen Replication

The study population comprised of 3,028 patients with T2D identified from an on-going study of the Genetics of Diabetes Audit and Research Tayside (Go-DARTS) and recruited in Tayside, Scotland, between 1st October 1997 and 1st March 2010. The study cohort was described in detail in earlier. A subset of Go-DARTS genotyped on Affymetrix platform was used in the current study. The baseline clinical characteristics of the Go-DARTS subset included in these analyses are very similar to the baseline clinical characteristics of the remaining Go-DARTS collective (data not shown). Hence, the subset of patients used for this analysis is very representative of the entire Go-DARTS cohorts**.**

Calculations for eGFR were made using the Modification of Diet in Renal Disease (MDRD) formula[128], which requires age, sex, race, and creatinine. Association testing for the 16 SNPs with eGFR for baseline was performed using linear regression analysis implemented in GPLINK[151] adjusting for age, sex, BMI, population structure, HbA1c, duration of diabetes, and systolic blood pressure. To investigate if the association of these loci with eGFR differed by albuminuria status, we conducted a stratified analysis with a test for interaction in patients with sustained normoalbuminuria (ACR<2.5mg/mmol in males and <3.5mg/mmol in females and duration of diabetes; >15 years at the end of follow-up) and in those with any albuminuria (ACR≥2.5mg/mmol in males and ≥3.5mg/mmol in females; either baseline or end of follow-up). Interaction of individual SNPs with albuminuria was tested using PLINK option 'interaction' with age, sex, BMI, albuminuria, and genotypes as covariates in the linear regression model.

To investigate if any of these polymorphisms were associated with rapid decline in renal function over the follow-up period, we performed a time to stage 3B of CKD (eGFR<45 mls/min/1.73m2) analysis. By using this cut-off, we lost 4% of our patients from the analysis (4% were prevalent cases (baseline eGFR<45 or Stage 3B CKD) therefore excluded from the analysis). If we had chosen to study progression to Stage 3A CKD (eGFR<60) we would have lost 20% patients from the analysis. Individuals with stage 3B CKD at baseline were excluded. Stage 3B of CKD was defined as three consecutive measurements of eGFR <45 mls/min/1.73m2 at least one month apart. Those who did not progress to stage 3B CKD were censored at the end of the follow-up period or date of death where applicable. We used a cox-proportional hazards model (proc phreg in SAS) with date of birth as 'time in'; 'last date' was defined as the first date of eGFR <45 mls/min/1.73m2 or the end of

follow-up period or date of death. Additional covariates included genotype, age, sex, BMI, and baseline eGFR.

We examined the association of these SNPS with albuminuria in the study population. 'Albuminuria case' was defined as ACR≥ 2.5 in males and ≥ 3.5 in females on their last ACR measurement and 'Albuminuria control' was defined as ACR<2.5 in males and < 3.5 in females and used in logistic regression analysis adjusting for age, sex and BMI.

We adopted a conservative threshold for significance (0.05/number of loci tested) and p-values below 0.003 were considered significant. All analyses were performed in PLINK version 1.07[151] and SAS 9.2. The study complies with the Declaration of Helsinki.

Power calculation

Power calculation for quantitative trait was completed with R (http://genome.sph.umich.edu/wiki/Power_Calculations:_Quantitative_Traits). For the quantitative trait, eGFR, the study had 99% power (at 5% level of significance) to detect an association with a SNP explaining 1% variability in the eGFR and 97% power to detect an association with a SNP explaining 0.5% variability in eGFR.

Genetic Risk Score

A weighted genetic risk score analysis (wGRS) was performed to test the joint effect of these 16 loci on baseline eGFR and time to CKD Stage 3B. We calculated wGRS (number of risk alleles*Beta), for each individual, using all 16 SNPs included in this study and tested the association of this GRS with baseline eGFR and time to CKD Stage 3B, adjusting for age, sex, BMI, hba1c, duration of diabetes, and systolic

blood pressure. The genetic risk score was also calculated for ~80 SNPS for BP and ~25 SNPs for HbA1C identified from the NGHRI database.

Model selection

Several epidemiological studies have shown that age, gender, duration of diabetes, BMI, (all at baseline), are important predictors of DN. For example, in the DCCT Type 1 diabetes study[152], time to diabetic nephropathy was related to age, AER, duration of diabetes, BMI, HbA1c (all at baseline), and gender. By adjusting for these same covariates in our 'maximum adjusted model', we aimed to identify the genetic variants which contribute to the variance in eGFR beyond the effect of these covariates. We also looked at the 'minimum adjusted model', with only age and gender as covariates in order to identify variants which can affect the eGFR by their effect on BMI and HbA1c. Since we did not see any difference in the two models, we present the results of maximum adjusted model only.

XX Specific methods for SUMMIT GWAS analysis for DKD

Phenotype Definitions

For the main phenotypes were patients with Type 1 diabetes and Type 2 diabetes with European ancestry. Albuminuria is classified based on timed overnight urinary albumin excretion rate (AER, µg/ min or mg/24 h) or an albumin-creatinine ratio (ACR, mg/mmol) as follows or patients with CKD and ESRD.

Table 11: Cut-off used for Phenotype definitions in SUMMIT GWAS analysis of DKD

	AER/min	AER 24 hours (UP)	ACR mg/mmol
Normoalbuminuria	<20µg	<30 mg	<2.5mg/mmol(M) <3.5mg/mmol(F)
High Microalbuminuria	100-200µg	150-300mg	12.5-25 mg/mmol(M) 17.5-35 mg/mmol(F)
Macroalbuminuria	>200 µg	>300 mg	>25 mg/mmol(M) >35 mg/mmol(F)
ESRD	eGFR<15ml/min or dialysis or kidney transplant		
Chronic Kidney Disease	eGFR is <60 ml/min		

Binary phenotypes

Diabetic albuminuria (DA):

- Cases: patients with any type of albuminuria or ESRD
- Controls: normo-albuminuria patients (duration of diabetes 15 years for T1D and 10 years for T2D)

Macroalbuminuria

- Cases: patients with macro-albuminuria or ESRD
- Controls: patients with Normo-albuminuria (duration of diabetes 15 years for T1D and 10 years for T2D)

Microalbuminuria

- Cases: patients with micro-albuminuria and high micro-albuminuria
- Controls: patients with Normo-albuminuria (duration of diabetes 15 years for T1D and 10 years for T2D)

Chronic kidney disease (CKD)

- Cases: patients with eGFR<60
- Controls: patients with eGFR>60 (duration of diabetes 15 years for T1D and 10 years for T2D)

Continuous phenotype

- eGFR: Log Transformed eGFR. Average of last two eGFR reading was taken.

Minimal adjustment model (min): Adjust model for age, duration of T2D, sex and Principle components from eigenstrat

Imputation

Imputation was performed based on HapMap2 NCBI build 36 reference haplotypes for first stage analysis and 1000 genomes reference haplotypes for the second stage analysis using SHAPTEit for phasing and IMPUTE 2 for imputation.

Analysis details

The primary analysis was a case control study that groups all individuals with microalbuminuria, high microalbuminuria, macroalbumiuria or ESRD as cases. Normoalbuminuric individuals with duration >15 years will be used as controls in Type 1 diabetes and Normoalbuminuric individuals with duration >15 years will be used as controls in Type 2 diabetes. Two regression models were used: one including only sex, duration, age at onset of diabetes and study specific covariates (min model), and one including HbA1c (max model). Duration of diabetes was defined as date of onset of diabetes to the date at which eGFR or ACR was measured in the lab. Thus, the duration of diabetes varied across different DKD phenotypes. The date of onset of diabetes had been generated in the Go-darts dataset and it was

based on an algorithm, which mainly used information about factors such as usage of diabetes drugs, HbA1c and the reported date of diagnosis on medical records.

SNPTEST Runs

SNPtest is installed on the linux server and called in a bash script. It is run sequentially on each of the 22 chromosomes. SNPtest takes in a genotype file of probabilities generated by IMPUTE 2 and takes a phenotype file. The order of id's in the phenotype and genotype files should match each other. SNPtest runs classical regression models and allows us to include covariates in the model. The -method option controls the way genotype uncertainty is taken into account when carrying out association tests. R script for doing quality control following SNPTEST run is in appendix 1.

Meta-analysis

Meta-analyses were implemented in GWAMA. Combined effect estimates were obtained using a fixed effect model. The Meta-analysis was performed using Chr:Position to identify SNPs not rs-ids to account for the different dbSNP .

Power Calculation for SUMMIT DKD studies

Power calculation for SUMMIT T1D and T2D were performed using CaTS[153]. Detailed results of power calculation are shown in Table 12 and Table 13. For the main DKD phenotype the SUMMIT studies have more than 75% power to detect a SNP with MAF of 10% and effect size (OR) of 1.4.

Table 12: Power calculations for T2D-DKD in SUMMIT studies

Phenotype	Case	Control	Power with MAF=10%	Power with MAF=20%
Diabetic Nephropathy	3,345	2,372	75	100
CKD	3,094	2,906	81	100
Micro-albuminuria	1,989	2,238	57	83
Macro-albuminuria	1339	2372	59	98
ESRD	371	2,076	5	14
Macroalbuminuria +CKD	897	1,610	8	47%
To detect a OR of 1.40 at alpha=0.00000001				

Table 13: Power calculations for T1D-DKD in SUMMIT studies

Phenotype	Case	Control	Power_10%MAF	Power_20%MAF
Diabetic Nephropathy	2,563	2,593	79	100
CKD	2353	1811	57	97
Micro-albuminuria	806	2,593	12	63
Macro-albuminuria	1,757	2,595	55	97
ESRD	813	2398	9	52
Macroalbuminuria+CKD	1,750	1,385	30	82
To detect a OR of 1.40 at alpha=0.00000001				

Results

I. Estimating Heritability of Various DN Phenotypes

The chip-based heritability of the several DKD phenotypes (based on albuminuria status) was estimated using GCTA. For this analysis, 4.5 million SNPs imputed with 1,000 genomes data for both the affymetrix and illumina cohort were merged using GTOOL. These SNPs had an information content of >0.75% and minor allele frequency of >1%.

Table 11 shows baseline characteristics of Go-DARTS cohort genotyped separately on the affymetrix and the illumina platform. The demographic characteristics were comparable in both the datasets expect age; those genotyped in the illumina cohort were older at baseline. The affymetrix dataset had 46% females, with a BMI of 30.6(5.3), eGFR of 73.9(18.7), HbA1c of 7.54(1.3) and an 8.71(7.4) year duration of diabetes at baseline. The illumina group had 42% females, with a BMI of 31.5(6.1), eGFR of 70.9(15.8), HbA1c of 7.3(1.4) and a 7.75(6.61) year duration of diabetes at baseline.

Table 12 shows the estimated chip-based heritability for various diabetic kidney disease phenotypes. The heritability estimates were highest for Macro-albuminuria &ESRD (macro-albuminuria and ESRD vs controls) phenotype followed by CKD+DN phenotype (highmicro&macro&eGFR<45 vs No Albuminuria &eGfR>60). The heritability estimates were lowest for micro-albuminuria about 1%. Continuous phenotypes eGFR and ACR had heritability of about 10-12%. Statistical significance was determined using the likelihood-ratio test of specific hypothesis. eGFR as a quantitative trait showed highest statistical significance (P value = 1.03E-05) followed by CKD phenotypes (P-value = 0.006). All the

remaining DKD phenotypes, except DN and micro-albuminuria showed statistically significant heritability at P-value threshold of 0.05.

Table 13 shows bivariate analysis of age and sex standardized residuals of average eGFR and average blood pressure during the study period. The study shows moderate genetic correlation between the two states 0.14(±0.36). These point estimates for r_g suggest that a small amount of the shared genetic determinants for eGFR and BP. The heritability estimates for blood pressure was 0.04(±0.03).

Table 14 shows bivariate analysis of age and sex standardized residuals of average eGFR and average HbA1c during the study period. The study shows high genetic correlation between the two states 0.71(±0.96) however, with a large SE. The heritability estimates for HbA1c was 0.02(±0.03) and lower than expected suggesting inadequate power of the study to estimate heritability and to do the bivariate analysis.

Table 14: Demographic Characteristics of the Go-DARTS Population

Characteristic, mean (sd)	Go-DARTS Affymetrix	Go-DARTS Illumina
Age at baseline, years	59.1 (11.0)	66.2 (11.6)
Sex,% female	46.4	42.3
Baseline BMI	30.6 (5.3)	31.5 (6.1)
Baseline eGFR, ml/min/1.73m^2	73.9 (18.7)	70.9 (15.8)
Baseline systolic blood pressure, mmHg	142.8 (18.4)	141.7 (18.8)
Baseline HbA_{1C}, mmol/mol	7.54 (1.3)	7.3 (1.4)
Baseline cholesterol, mmol/L	4.40 (0.97)	4.34 (0.91)
Duration of diabetes at baseline, years	8.71 (7.44)	7.75 (6.61)

Table 15: Narrow –Sense Heritability of DN Phenotypes using GCTA

Phenotype	narrow sense Heritability	SE	P-value
DN(any albuminuria Vs No Albumiuria)	0.08	0.07	0.1
Micro(micro-albuminuria Vs No Albumiuria)	0.01	0.07	0.49
Macro+ESRD(macroalbuminuria+ESRD Vs No Albumiuria)	0.21	0.11	0.03
ESRD(ESRD Vs No Albumiuria)	0.23	0.13	0.03
ESRD Vs ALL (ESRD Vs ALL)	0.08	0.06	0.06
CKD(eGFR<60 vs eGFR>60+10 year duration of diabetes)	0.12	0.05	0.006
eGFR(quantitative trait)	0.12	0.03	1.03E-05
CKD-DN(highmicro+macro+eGFR<45 vs No Albumiuria +eGfR>60	0.21	0.14	0.04
ACR	0.09	0.05	0.04
sRAGE	0.37	0.19	0.01
Systolic blood Pressure	0.04	0.03	0.1
eGFR	0.12	0.03	0.04

Table 16: Bivariate analysis of age and sex standardized residuals of average eGFR and average systolic blood pressure during the study period

	Point Estimate	SE
eGFR	0.12	0.03
Systolic Blood Pressure	0.04	0.03
r_g	0.14	0.36

Table 17: Bivariate analysis of age and sex standardized residuals of average eGFR and average HbA1c during the study period

	Point Estimate	SE
eGFR	0.12	0.03
HbA1c	0.02	0.03
r_g	0.71	0.96

I. Association of known SNPs for upstream risk factors for DN and their association with Diabetic Kidney Disease

Table 12 shows baseline characteristics of Go-DARTS cohort genotyped separately on the affymetrix and the illumina platform. The demographic characteristics were comparable in both the datasets expect age; those genotyped in the illumina cohort were older at baseline. The affymetrix dataset had 46% females, with a BMI of 30.6(5.3), eGFR of 73.9(18.7), HbA1c of 7.54(1.3) and an 8.71(7.4) year duration of diabetes at baseline. The illumina group had 42% females, with a BMI of 31.5(6.1), eGFR of 70.9(15.8), HbA1c of 7.3(1.4) and a 7.75(6.61) year duration of diabetes at baseline. Genotype data for the candidate SNPs for BP and HBA1C in the affymetrix and illumine platforms were merged using GTOOL.

We utilized the NHGRI Catalogue of Published Genome-Wide Association Studies (updated July 2013) to identify gene/loci associated with BP (n=60) and HbA1c (n=25) with P-value <10-15 and at least one independent replication. For the quantitative traits, BP and HbA1C, the study had more than 90% power (at 5% level of significance) to detect an association with a SNP explaining 1% variability in the BP and HbA1C. Genotype distribution for these markers was in Hardy-Weinberg equilibrium (P >0.001) and the MAF was comparable to HAPMAP CEPH. Given the strong prior information about the role of the SNP tested here, we considered this a replication study, and therefore, p values below 0.05 could be considered significant. However, we adopted a more conservative threshold for significance (0.05/number of loci tested) and p values below 0.003 were considered significant.

Table 15 shows replication results of SNPs associated with BP in Type 2 diabetes Go-DARTS with P-value <0.001). SNPs in *MTHFR, CAPZA1, GPR39, CNTN4, ADH1C-ADH7, SLC39A8, C5orf23-LOC340113, RANBP3L-SLC1A3, TRPA1-LOC100129527, MAL2-NOV, ZFAT, LOC100128248-LOC100131290, TBX3-LOC100129020,TBX3-LOC100129020, GRINL1A-ALDH1A2, CSK,LOC100132798, CDH13,* and *FAT1P1-C20orf187* were significantly associated with BP after correction for multiple testing. The direction of effect of these SNPs was consistent with the direction of effects reported in the original GWAS papers.

Table 16 shows replication results of SNPs associated with HbA1c in Type 2 diabetes Go-DARTS with P-value <0.01). SNPs in *G6PC2, ABCB11, ANK1, ANK1, HK1, TCF7L2, FN3KRP* were significantly associated with HbA1c after correction for multiple testing. The direction of effect of these SNPs was consistent with the direction of effects reported in the original GWAS papers.

Table 17 shows association of beta-weighted genetic risk scores for BP (N=60) and HbA1c (n=25). In a logistic regression model with age and sex as covariates, the genetic risk scores for both BP (OR=25 CL=3.44-166, P=0.001) and HbA1c (OR=1.26 CL=1.16-1.38, P-value<0.0001) were significantly associated with the likelihood of having macro-albuminuria and ESRD.

Table 18 : Replication of SNPs associated with BP in Type 2 diabetes Go-DARTS (n=7276) (Total SNPS~ 60)

CHR	SNP	BP	A1	BETA	P	Gene
1	rs17367504	11862778	G	-2.067	4.59E-04	*MTHFR*
1	rs17030613	1.13E+08	C	1.298	2.92E-03	*CAPZA1*
2	rs13420028	1.33E+08	G	-1.858	4.42E-04	*GPR39*
3	rs4370013	2654691	T	-2.005	9.52E-04	*CNTN4*
4	rs991316	1E+08	C	-0.8756	7.82E-03	*ADH1C-ADH7*
4	rs13107325	1.03E+08	T	-1.738	6.49E-03	*SLC39A8*
5	rs1173771	32815028	G	-0.8925	4.11E-03	*C5orf23-LOC340113*
5	rs7735940	36423931	T	-0.8506	2.01E-03	*RANBP3L-SLC1A3*
8	rs1963982	73106916	G	-1.191	1.02E-03	*TRPA1-LOC100129527*
8	rs2469997	1.2E+08	C	-0.8941	9.53E-04	*MAL2-NOV*
8	rs7827545	1.36E+08	T	-1.663	6.87E-05	*ZFAT*
12	rs7960483	45925755	T	-1.046	3.55E-04	*LOC100128248-LOC100131290*
12	rs2384550	1.15E+08	A	-1.196	3.18E-03	*TBX3-LOC100129020*
12	rs35444	1.16E+08	G	-1.203	2.24E-03	*TBX3-LOC100129020*
15	rs1550576	58213414	T	-1.701	5.63E-03	*GRINL1A-ALDH1A2*
15	rs1378942	75077367	A	-0.875	2.76E-03	*CSK*
15	rs2398162	96830550	G	-1.473	3.77E-03	*LOC100132798*
16	rs3096277	83764204	C	-0.736	6.73E-03	*CDH13*
20	rs1327235	10969030	G	-1.028	2.78E-03	*FAT1P1-C20orf187*

Table 19: Replication of SNPs associated with HbA1c in Type 2 diabetes Go-DARTS (n=7276) (Total SNPS~ 25)

CHR	SNP	BP	A1	BETA	P	gene
2	rs1402837	1.7E+08	T	-0.07404	1.33E-02	*G6PC2*
2	rs552976	1.7E+08	G	-0.08739	6.87E-05	*ABCB11*
8	rs6474359	41549194	C	0.3232	9.27E-03	*ANK1*
8	rs4737009	41630405	A	-0.081	2.85E-02	*ANK1*
10	rs16926246	71093392	T	-0.08038	4.01E-02	*HK1*
10	rs7072268	71099913	C	-0.1065	1.80E-05	*HK1*
10	rs7903146	1.15E+08	T	0.06434	1.67E-02	*TCF7L2*
17	rs1046896	80685533	T	-0.1029	1.08E-03	*FN3KRP*

Table 20: Genetic risk scores for BP (~60 SNPs), HbA1C (~25 SNPs) and its association with DKD

Phenotype	OR(95%CI) for wGRS for BP*	P-value for association with wGRS for BP	OR(95%CI) wGRS for HbA1c**	P-value for association with wGRS for HbA1c
Any Albuminuria	4 (1.28-12.5)	0.01	1.08 (1.02-1.03)	0.007
Microalbuminuria	3.03 (0.9-13.09)	0.06	1.03 (0.9-1.08)	0.33
Macro-albuminuria +ESRD	25 (3.44-166)	0.001	1.26 (1.16-1.38)	<0.0001

*adjusted for age, sex and baseline HbA1c
**adjusted for age, sex and baseline BP

III. GWAS for biomarker for DN (sRAGE)

RESULTS

Table 18 shows demographic characteristics of the study participants. Genotype data and sRAGE was available in 589 participants in CARDS, 348 in Go-DARTS-Affymetrix and 440 in Go-DARTS-Illumina Datasets. The demographic characteristics in the three datasets were similar expect duration of diabetes which was higher in the Go-DARTS dataset. The median (inter quartile range) sRAGE levels was 1426 pg/ml (1108-1812) in CARDS dataset and 1303.8 pg/ml (1017.5-1702.2) in Go-DARTS dataset. Median (inter quartile range) esRAGE levels in the CARDS dataset were 340 pg/mL (250–460). The sRAGE and esRAGE levels in CARDS dataset were highly correlated, with a Pearson's correlation coefficient of 0.88 ($P<0.0001$). **Figures 3 and 4** shows distribution of log transformed sRAGE levels in CARDS and Go-Darts datasets.

Figure 5 shows a Q-Q plot of the –log10 p-values from a meta-analysis of genome-wide association study for sRAGE levels in CARDS and Go-DARTS data set; the cumulative distribution of test-statistic follows the null distribution over most of its range but there is a tail of extreme results. Most of these extreme results arise from chromosome 6 as seen in the Manhattan plot (**Figure 6**) of the genome-wide association between the SNPs and sRAGE levels. A similar distribution of P-values on Q-Q plots was seen in the CARDS, Go-DARTS Affymetrix and Go-DARTS-Illumina datasets (**Figures 8, 9, 10**).

Table 19 shows results of SNP association tests of the top 22 SNPS (with $P \leq 10E\text{-}8$) with total sRAGE levels in CARDS-Go-DARTS meta-analysis. The top markers

were contained in the genes *PPT2, LY6GD6, AGER, NOTCH4* and *HLA-DQA-2* and *HLA-DQB1* region. The estimates of effect (β) are for the log-transformed response variable. rs2070600 (G28S) in the *AGER* gene was the topmost marker associated with sRAGE levels with the 'T' allele associated with lower sRAGE levels (β =-0.14, P=8.25E-18).

Using the summary level data from CARDS-Go-DARTS meta-analysis, we then performed a genome-wide conditional analysis implemented in GCTA to identify markers independently associated with sRAGE levels across the genome (**Table 20**). Our genome-wide conditional analysis showed presence of two independent loci in *AGER-HLA-DQA1* gene regions, rs2070600 (G28S) (β = -0.13, P=4.32E-13) and rs9272346 (β = 0.04, P=2.40E-07) (**Figure 8**). Three additional markers, rs6857222 in the *COX5BLI-LOC441026* gene (BETA= -0.05 P=8.76E-08), rs2807326 on the WNT4-*LOC343384* gene region (BETA= 0.04 P=3.47E-07), and rs10940285 in the *ITGA1-ITGA2* gene region (BETA= 0.04 P=1.85E-07) were independently associated with sRAGE levels.

Table 21 shows the association of the 5 SNPs from the genome-wide conditional analysis (with P≤10-E6) in 666 patients with Type 2 diabetes in SDR study, 140 patients with Type 2 diabetes in the KORA study and 3,456 patients with Type 1 diabetes in the FinnDiane Study. The association of two markers in the *AGER-HLA* region (rs2070600 and rs9272346) replicated in the Type 1 and Type 2 diabetes cohorts while rs10940285 replicated in Type 2 diabetes cohort and was borderline significant in Type 1 diabetes cohort. The direction of effect for these three SNP was consistent in all the cohorts. rs6857222 in *COX5BLI-LOC441026* gene and rs2807326 in *WNT4-LOC343384* gene region did not replicate in Type 1 or Type 2 diabetes cohorts.

Using summary levels statistics in CARDS and Go-DARTS cohorts, we estimated chip based heritability of sRAGE levels. GCTA showed a point estimate of h^2 = 0.37 for circulating sRAGE levels (standard deviation (sd) = 0.19, P-value=0.01) consistent with high heritability. The p-values are from likelihood tests with null hypothesis of heritability being zero. After adjustments for age, sex and BMI, three SNPs (rs10940285, rs2070600, and rs9272346) explained 10% variation (adjusted r-squared in SAS) in circulating sRAGE levels in the SDR dataset (validation cohort). There was no LD between *AGER* SNP rs2070600 (*G28S*) and *HLA-DQA1* SNP rs9272346 in the SDR cohort (r^2=0.02) confirming that rs9272346 is not directly tagging rs2070600 in *AGER* gene.

We looked at the association of SNPs with total esRAGE levels with $P \leq 10 \times 10^{-8}$ in the CARDS dataset. The associations with SNPs showed similar trend as association with sRAGE levels with a slightly lower order of magnitude. We did not identify any variants differentially affecting esRAGE and sRAGE and esRAGE/sRAGE ratio levels in the CARDS dataset.

Table 21: Demographics and Baseline Characteristics of the Three Datasets

	CARDS (n=589)	Go-DARTS Affymetrix (n=348)	Go-DARTS Illumina (n=440)
Age	61.8(±7.9)	56(±12.6)	71.6(±10.9)
Sex(% Female)	30%	38%	41%
BMI	30.9(±5.02)	30.9(±5.07)	31.3(±6.07)
Duration of Diabetes	7.89(±6.2)	9.8(±7.6)	13.1(±6.7)
HbA1c	7.87(±1·42)	7.5(±1.26)	7.4(±1.40)
sRAGE levels(log10 transformed)	3.13(±0.18)	3.11(±0.18)	3.12(±0.18)
Platform	Perlegen 5	Affymetrix	Illumina 6
Sample Call Rate	>98%	>98%	>98%
pHWE exclusion	10^{-5}	10^{-5}	10^{-5}
Reference Panel/imputation software	HAPMAP 2 / IMPUTE2	HAPMAP 2/IMPUTE2	HAPMAP 2 / IMPUTE2
SNPs after imputation and QC	2,298,008	2,477,804	2,351,765

Table 22: Distribution of raw sRAGE values in CARDS and GO-DARTS datasets

	Distribution of raw sRAGE values in CARDS and Godarts			
	Mean	Median	Maximum	Minimum
CARDS	1501.67	1426	4626	169.99
GAODARTS	1451.68	1303.80	4955.9	271

Figure 3: Distribution of Log10-sRAGE in CARDS

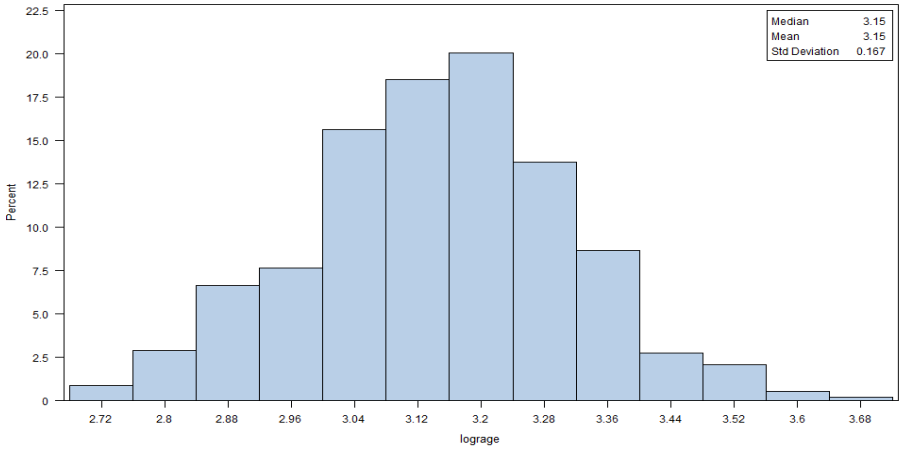


Figure 4: Distribution of Log10-sRAGE in Go-DARTS dataset

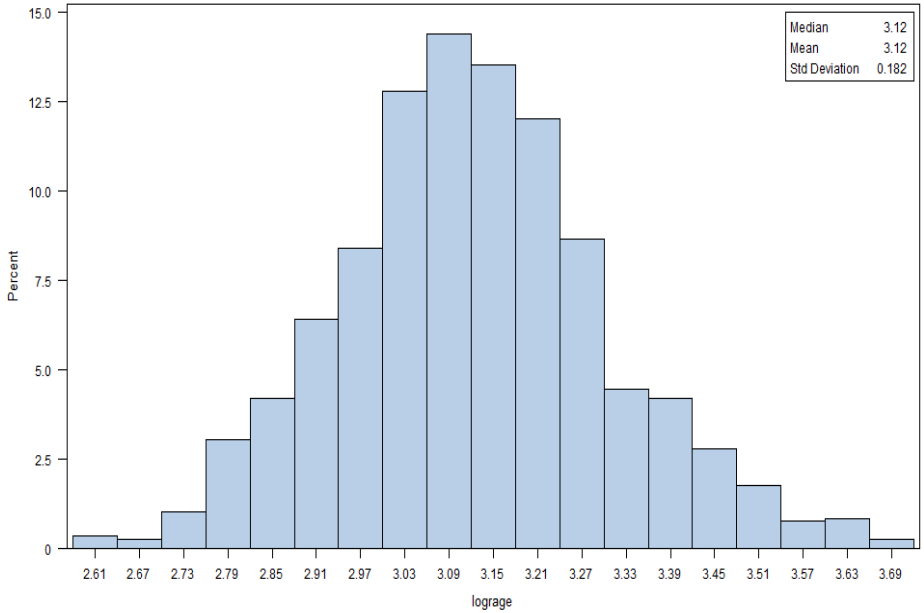


Table 23: GWAS Meta-analysis for sRAGE in CARDS and GO-DARTS cohorts

					Cards_perlegen			Godarts_illumina			Godarts_Affy			Meta-analysis			
CHR	**BP**	**MARKER**	**EA**	**MAF_average**	**BETA**	**SE**	**Pval**	**beta**	**SE**	**Pval**	**beta**	**SE**	**Pval**	**beta**	**SE**	**Pval**	**Gene**
6	32259421	rs2070600	T	0.08	-0.17	0.02	4.99E-12	-0.16	0.05	8.91E-04	-0.1	0.03	4.91E-05	-0.14	0.02	8.25E-18	PBX2 SRAGE
6	32232402	rs10947233	T	0.08	-0.16	0.02	7.99E-13	NA	NA	NA	-0.11	0.02	1.89E-05	-0.14	0.02	8.57E-17	PPT2
6	32293583	rs2854050	G	0.08	-0.13	0.02	2.82E-11	0.17	0.05	3.65E-04	0.08	0.02	5.38E-04	-0.12	0.01	2.64E-16	NOTCH4
6	31790008	rs805284	G	0.08	-0.14	0.02	8.40E-12	NA	NA	NA	0.09	0.02	3.05E-04	-0.12	0.02	1.33E-14	LY6G6D
6	32930164	rs4148878	T	0.08	-0.14	0.02	1.49E-10	0.05	0.04	2.64E-01	0.08	0.03	2.81E-03	-0.1	0.02	5.17E-12	TAP1 PSMB9
6	32288409	rs2071285	T	0.09	-0.16	0.03	1.66E-09	NA	NA	NA	-0.09	0.03	4.05E-04	-0.12	0.02	8.77E-12	NOTCH4
6	32822121	rs2071800	G	0.08	-0.1	0.02	2.77E-08	0.14	0.05	3.92E-03	0.07	0.03	8.07E-03	-0.09	0.01	2.19E-11	HLA-DQA2
6	32287410	rs2854047	T	0.09	-0.16	0.03	9.06E-09	NA	NA	NA	-0.09	0.03	3.82E-04	-0.12	0.02	5.63E-11	NOTCH4
6	32287472	rs2555469	G	0.09	-0.16	0.03	9.06E-09	NA	NA	NA	-0.09	0.03	3.86E-04	-0.12	0.02	5.72E-11	NOTCH4
6	32820796	rs17500510	G	0.09	NA	NA	NA	0.13	0.04	1.34E-03	0.06	0.02	9.18E-03	-0.08	0.01	6.84E-11	HLA-DQA2
6	32274362	rs8192575	G	0.09	-0.18	0.03	3.93E-08	NA	NA	NA	-0.09	0.03	2.53E-04	-0.13	0.02	2.20E-10	NOTCH4
6	32788554	rs7764819	T	0.14	-0.07	0.02	2.15E-06	0.07	0.02	1.07E-03	0.06	0.02	9.61E-03	-0.07	0.01	2.49E-10	HLA-DQB1
6	32765659	rs9405119	G	0.25	-0.06	0.01	6.02E-06	-0.05	0.02	1.34E-02	-0.05	0.02	1.04E-03	-0.05	0.01	9.21E-10	HLA-DQB1
6	32790790	rs9275602	C	0.22	-0.09	0.02	9.43E-07	0.08	0.03	4.87E-03	0.05	0.02	1.26E-02	-0.07	0.01	1.90E-09	HLA-DQB1
6	32771666	rs5000632	T	0.24	-0.05	0.01	1.41E-05	0.05	0.02	1.41E-02	0.05	0.02	1.06E-03	-0.05	0.01	2.25E-09	HLA-DQB1
6	32772949	rs9348891	G	0.24	-0.05	0.01	1.41E-05	-0.05	0.02	1.40E-02	-0.05	0.02	1.14E-03	-0.05	0.01	2.39E-09	HLA-DQB1
6	32773145	rs9394113	G	0.24	-0.05	0.01	1.41E-05	-0.05	0.02	1.40E-02	-0.05	0.02	1.15E-03	-0.05	0.01	2.40E-09	HLA-DQB1
6	32787710	rs9378125	G	0.29	-0.06	0.01	1.54E-05	-0.04	0.02	5.55E-02	-0.06	0.02	2.64E-04	-0.05	0.01	3.62E-09	HLA-DQB1
6	32712350	rs9272346	G	0.41	0.06	0.01	8.54E-07	0.02	0.01	1.60E-01	0.05	0.01	9.17E-04	0.05	0.01	7.30E-09	HLA-DQA1
6	32265342	rs2856437	G	0.08	-0.13	0.04	1.92E-03	0.17	0.05	3.65E-04	0.09	0.02	3.12E-04	-0.11	0.02	1.00E-08	GPSM3
6	31777475	rs1266071	T	0.12	-0.07	0.02	1.93E-05	-0.08	0.04	3.88E-02	-0.06	0.02	7.78E-03	-0.07	0.01	5.50E-08	BAT5
6	32734289	rs6906021	T	0.48	-0.05	0.01	1.21E-05	0.02	0.02	3.74E-01	0.05	0.01	1.33E-03	-0.04	0.01	6.60E-08	HLA-DQB1

Table 24 Genome-wide conditional analysis of CARDS-Go-DARTS dataset

CARDS GODARTS meta-analysis results						Joint analysis results			Gene
Chr	SNP	freq	b	se	p	bJ	bJ_se	pJ	
1	rs2807326	0.37	0.04	0.01	5.30E-07	0.04	0.01	3.47E-07	WNT4-LOC343384
4	rs6857222	0.14	-0.05	0.01	1.86E-05	-0.05	0.01	8.76E-08	COX5BLI-LOC441026
5	rs10940285	0.33	0.03	0.01	3.61E-06	0.04	0.01	1.85E-07	ITGA1-ITGA2
6	rs2070600	0.07	-0.14	0.02	8.25E-18	-0.13	0.02	4.32E-13	RAGE
6	rs9272346	0.39	0.05	0.01	7.30E-09	0.04	0.01	2.40E-07	RAGE-HLADQA1

Table 25 Replication of the top hits from CARDS- Go-DARTS meta-analysis in Type 1 and Type 2 Diabetes cohorts

CARDS + Go-DARTS meta-analysis (n=1373)					Replication in Type 2 diabetes cohorts SDR(n=666) and Kora (n=140)			Replication in Type 1 diabetes cohort Finndiane(n=3546)			Meta-analysis in type 2 Diabetes cohorts		Meta-analysis in Type1 and Type 2 Diabetes cohorts		
Chr	SNP	Beta	SE	P-value	Beta	SE	P-value	Beta	SE	P-value	z-score	P-value	z-score	P-value	Gene
1	rs2807326	0.04	0.01	3.47E-07	0.01	0.01	2.69E-01	0.01	0.01	0.59	3.42	6.34E-04	1.61	1.06E-01	WNT4-LOC343384
4	rs6857222	-0.05	0.01	8.76E-08	-0.01	0.01	4.47E-01	-0.01	0.01	0.23	4.66	3.11E-06	2.13	3.33E-02	COX5BLI-LOC441026
5	**rs10940285**	**0.04**	**0.01**	**1.85E-07**	**0.03**	**0.01**	**1.69E-03**	**0.01**	**0.01**	**0.05**	**6.06**	**1.33E-09**	**5.43**	**5.67E-08**	**ITGA1-ITGA2**
6	**rs2070600**	**-0.13**	**0.02**	**4.32E-13**	**-0.13**	**0.02**	**1.61E-11**	**-0.13**	**0.01**	**9.36E-54**	**-9.75**	**1.89E-22**	**-18.58**	**4.78E-77**	**RAGE**
6	**rs9272346**	**0.04**	**0.01**	**2.40E-07**	**0.02**	**0.01**	**3.63E-02**	**0.03**	**0.01**	**2.54E-05**	**5.34**	**9.40E-08**	**6.52**	**7.08E-11**	**RAGE-HLA-DQA1**

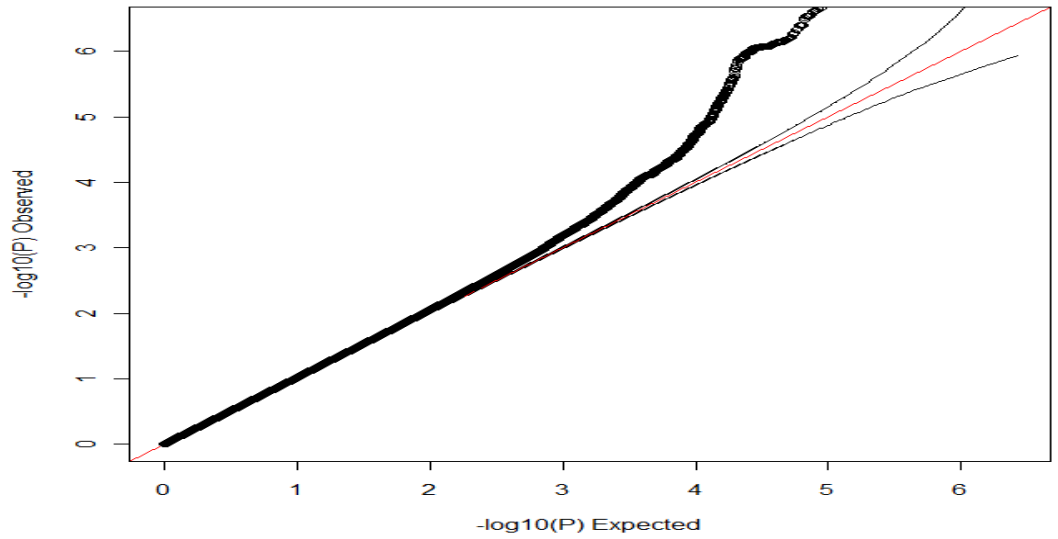


Figure 5: QQ plot of GWAS meta-analysis for sRAGE in CARDS and Go-DARTS dataset

Figure 6: Manhattan plot of GWAS meta-analysis

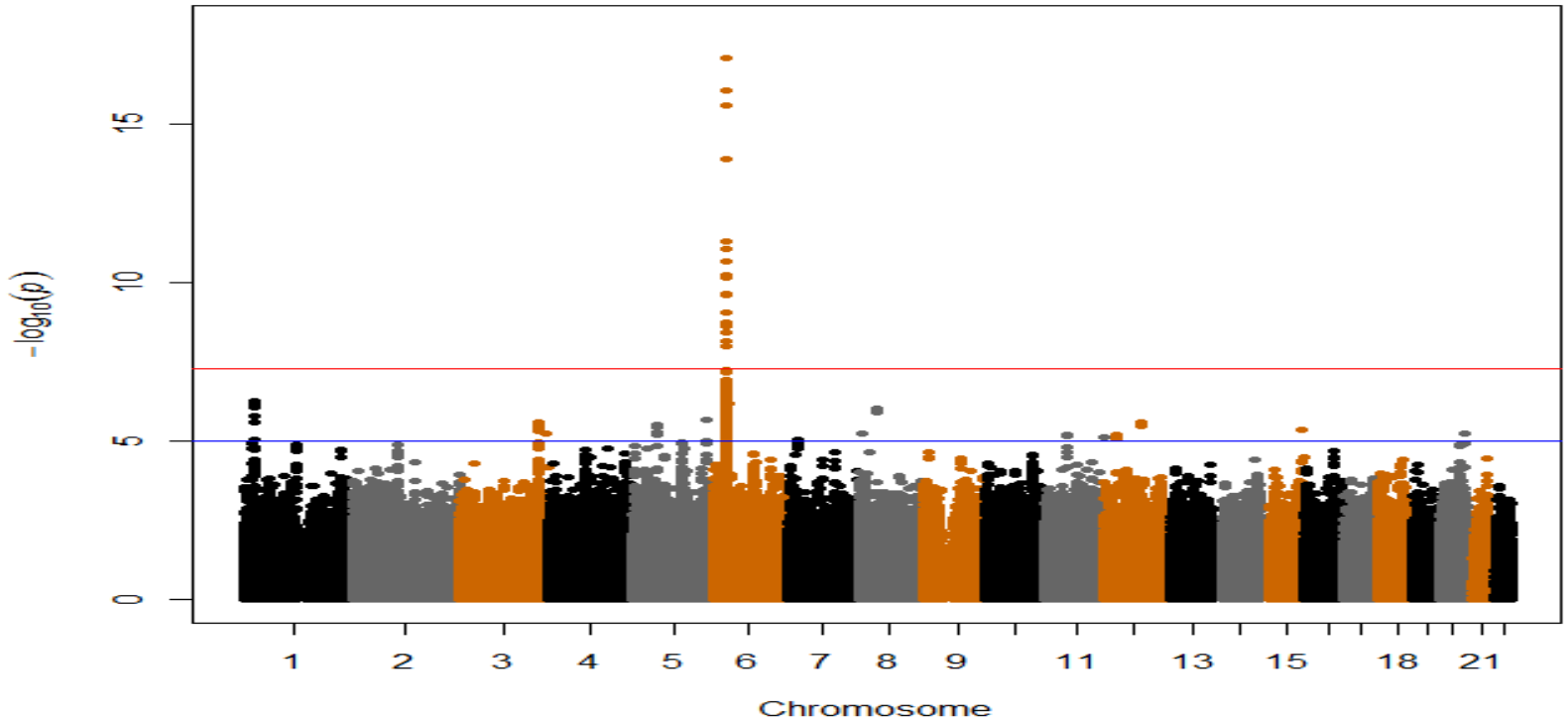


Figure 6: Manhattan plot of GWAS meta-analysis for sRAGE in CARDS and Go-DARTS dataset

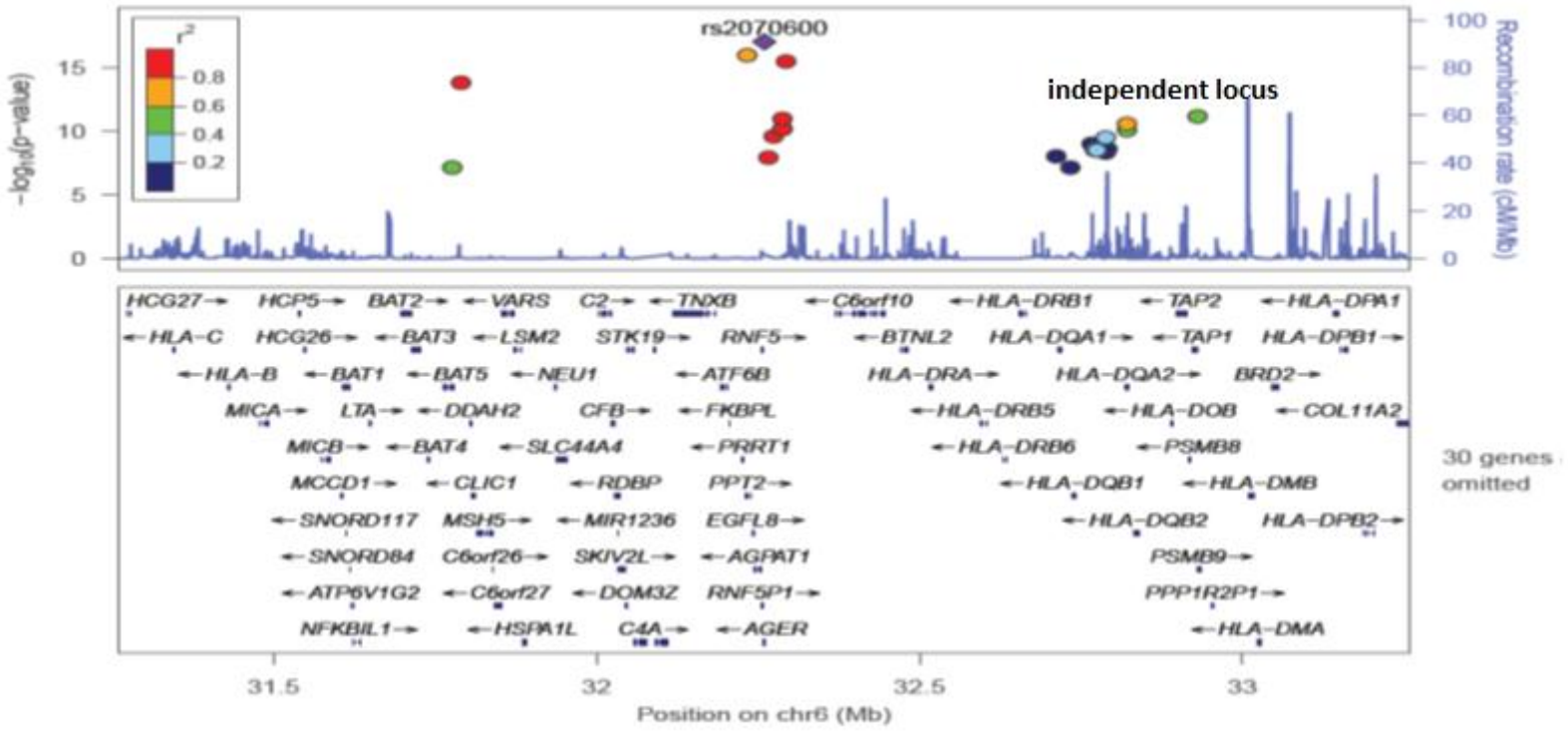


Figure 7: Regional Association plot RAGE-HLA gene region – association with circulating sRAGE levels

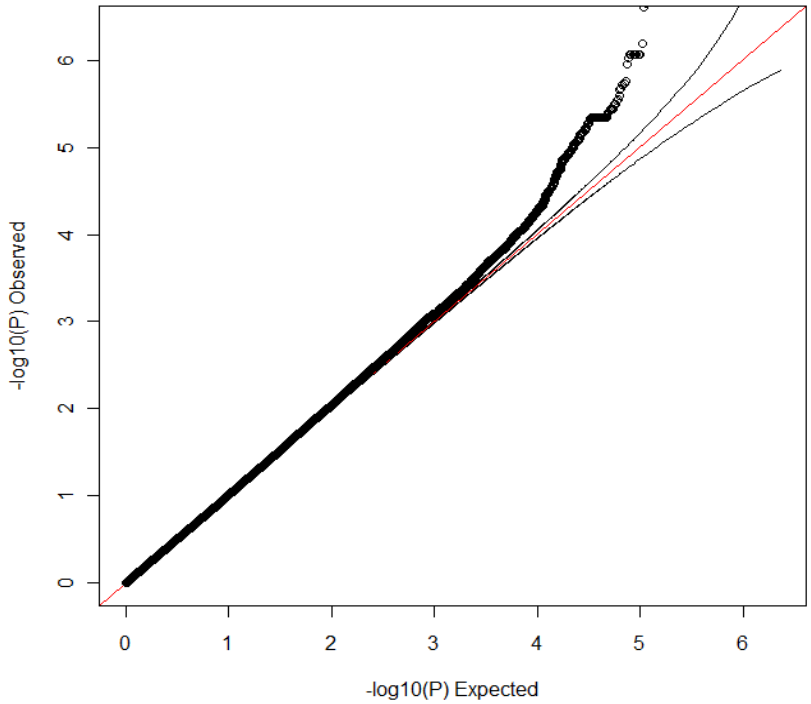


Figure 8: QQ plot for GWAS of sRAGE levels of CARDS Perlegen dataset

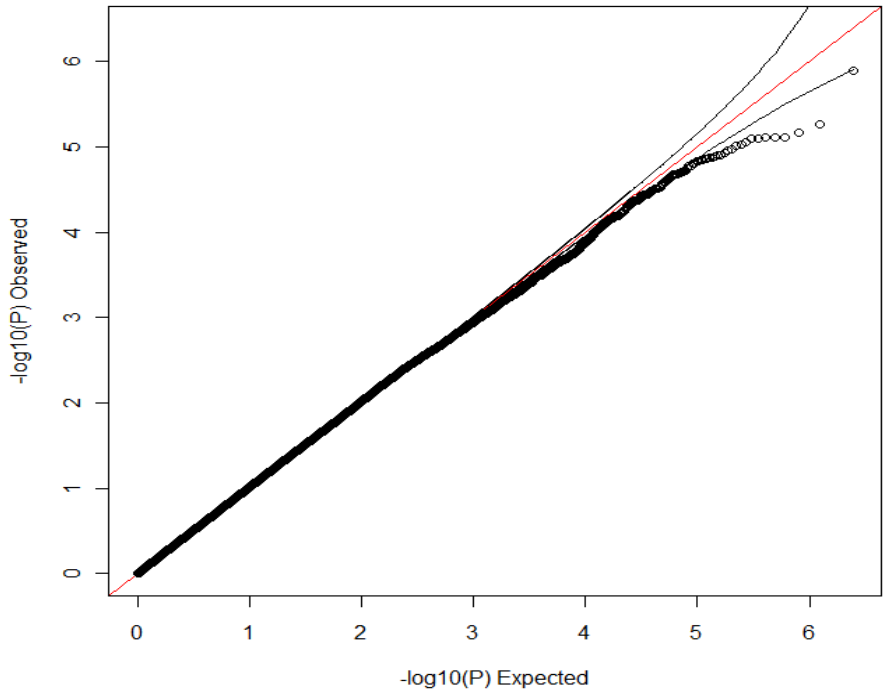


Figure 9: QQ plot for GWAS of sRAGE levels of Go-DARTS affymetrix dataset

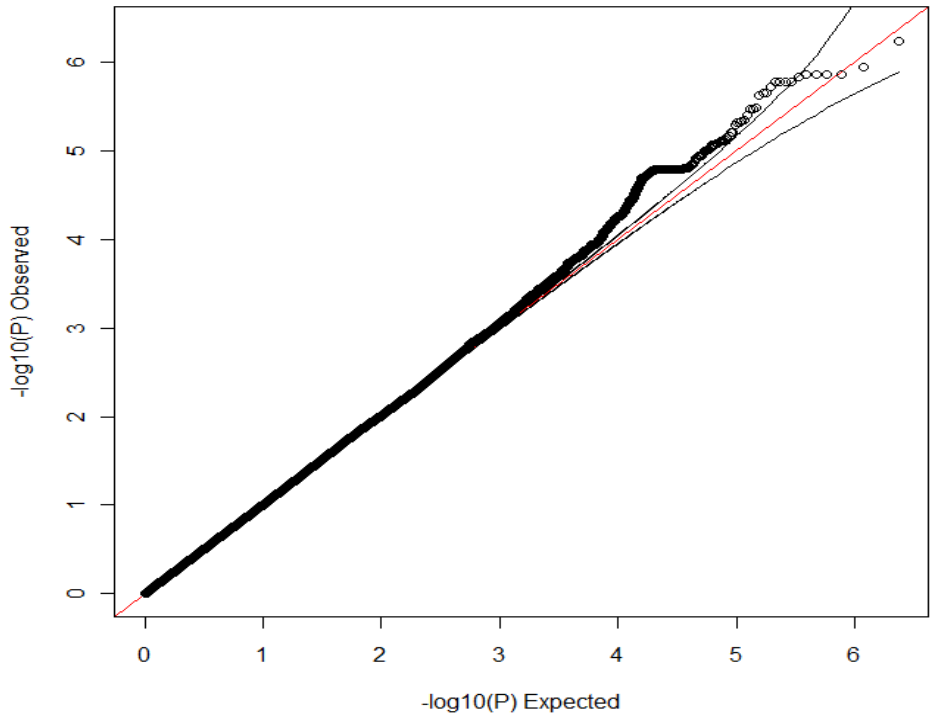


Figure 10: QQ plot for GWAS of sRAGE levels of Go-DARTS Illumina dataset

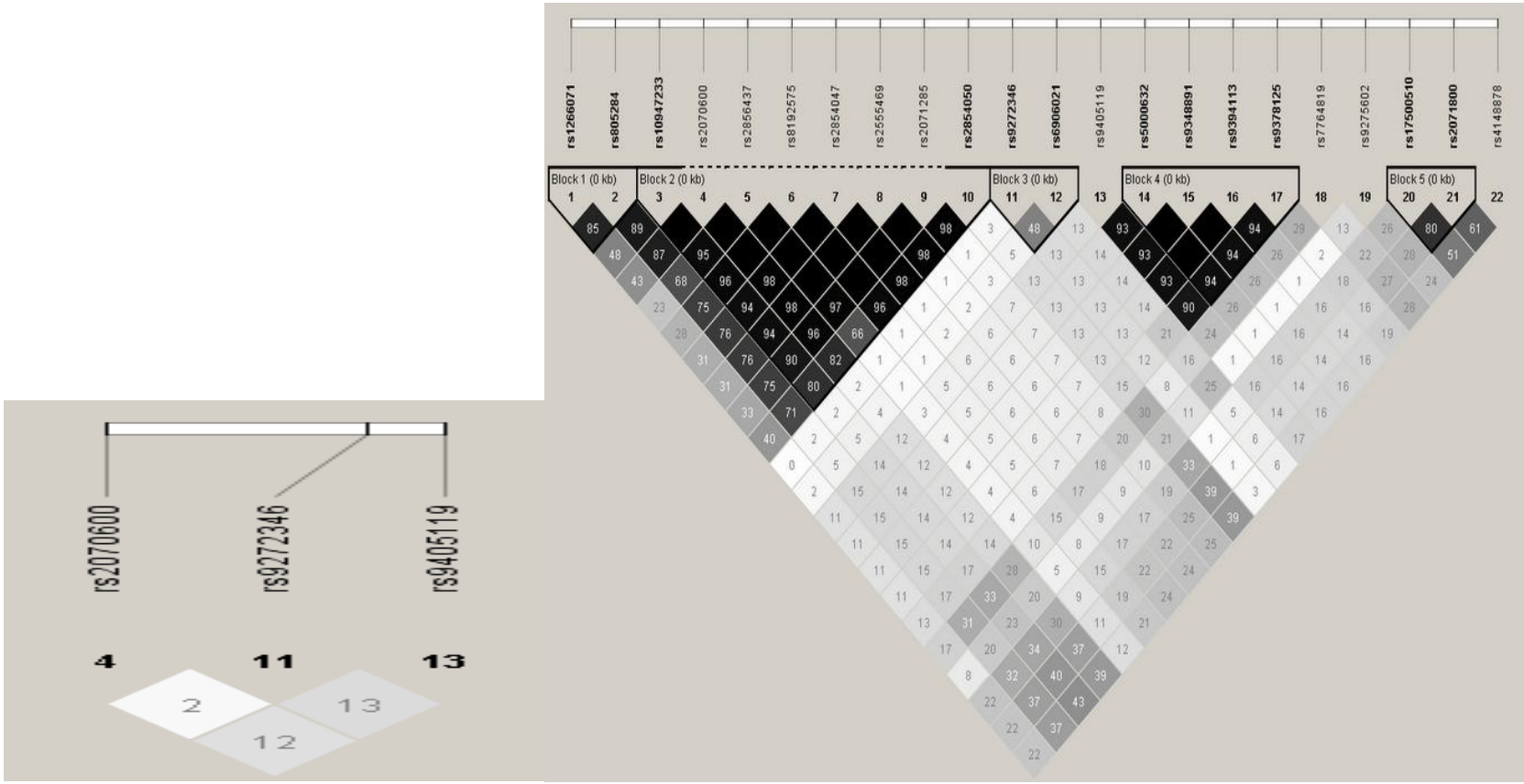


Figure 11: Result of LD based pruning of the top 22 markers associated with sRAGE levels in CARDS and Go-DARTS meta-analysis

IV. Replication of SNPs in CKDgen I consortium and contribution to CKDgen II consortium

Results 1: Replication of CKDgen Consortium I

Table 24 shows baseline characteristics of the study population. Genotype data was available for 3,028 patients (46.4% females) with T2D with mean baseline BMI of 30.6(±5.3), mean age 59.1(±11), mean HbA1c 58 mmol/mol (7.54(±1.3) and consisted of 57% males. The mean follow-up period for the entire study was 10.6(±9.1) years with a median of three eGFR readings/year/person (IQR=2-4) and a mean baseline eGFR of 73.9(±18.7) mls/min/1.73m^2. There were no significant difference in the LDLc levels in the individuals genotyped on affymetrix and illumina chip at baseline. In those genotyped on affymetrix platform the ldl-c levels were 2.03 (±0.93) mmol/l and on the illumina platform the LDLc levels were 2.11 (±0.98) mmol/l. Table 25 shows baseline characteristics of cases that progressed to Stage 3B CKD and controls. As expected, those who progressed to Stage 3B CKD were likely to be older, with a higher BMI, HbA1c, systolic BP and longer duration of diabetes

Figure 6 is the Linkage-Disequilibrium plot of the 16 SNPs in the study and shows all were in linkage equilibrium with each other. Table 26 shows the identified association of 16 eGFR-associated loci with baseline eGFR, with eGFR stratified by albuminuria status, and the association of these polymorphisms with time to Stage 3B CKD. The minor alleles 'T' of *GCKR* (Beta=1.30 P-value=3.23E-03), and 'T' of *UMOD* (Beta=2.0 P-value=8.84E-04) were associated with a higher eGFR at baseline; the minor allele 'A' of *SHROOM3* (Beta= -1.28 P-value=3.18E-03) was

associated with a lower eGFR at the predefined threshold (P≤0.003). None of the other SNPs was associated with baseline eGFR. None of the SNPs included in the study were associated with albuminuria after correction for multiple testing (Table 26). In patients with sustained normoalbuminuria (n=613), minor allele 'T' of *UMOD* was associated with higher eGFR (beta=4.03, P-value=1.10E-03) while, in patients with albuminuria (n=2,096), minor allele 'T' of *GCKR* (Beta=1.12, P-value=4.27E-02) and 'A' of *SHROOM3* (Beta=-1.43, P-value=7.28E-03) were associated with a lower eGFR. Of the 17 SNPs *UMOD* (HR= 0.83(0.70, 0.99), P-value=0.03), *PIP5K1B* (HR= 0.85(0.75, 0.96), P-value=0.01) and *SLC7A9* (HR= 0.86(0.76, 0.98) P-value=0.02) was associated with time to Stage 3B CKD (eGFR<45mls/min/1.73m2) at 0.05 threshold for significance. Since these are genetic variants are associated with age-related decline in eGFR in the general population (and not any disease specific decline), we used time to event analysis with date of birth as starting point. However, we performed a sensitivity analysis in which we used starting point as the baseline of Go-DARTS study. Although this analysis decreases power due to reduction in the person-year follow-up, we see similar effect size of association with progression to CKD Stage 3B. For example, the hazard ratio of *UMOD* with time to CKD Stage 3B with starting point of Go-DARTS study (HR= 0.87(0.74, 1.03) P-value=0.1) is very similar to the hazard ratio with date of birth as a starting point. The wGRS for the 16 SNPs explained 1% variation in baseline eGFR and was significantly associated with baseline eGFR after adjustments for age, sex, BMI, HbA1c, duration of diabetes and systolic blood pressure (P=0.0026 Beta= 0.84(±0.28). The wGRS was not associated with time to Stage 3B CKD (P=0.52). None of the markers were associated with albuminuria

after correction for multiple testing however; *SHROOM3* and *UMOD* were significant at 0.05 threshold of significance.

Table 26: Reported eGFR associated loci in Kottgen A et al paper

First Author	Region	Chr_id	Chr_pos	Reported Gene(s)	SNPs	p-value	Initial Sample Size	Replication Sample Size
Kottgen A	1q21.3	1	150951477	ANXA9,FAM63A,PRUNE,BNIPL,LASS2,SETDB1	rs267734	1.00E-12	67,093	22,982
Kottgen A	2p23.3	2	27730940	GCKR, IFT172, FNDC4	rs1260326	3.00E-14	67,093	22,982
Kottgen A	2p13.1	2	73868328	NAT8,NAT8B,ALMS1	rs13538	5.00E-14	67,093	22,982
Kottgen A	3q23	3	141807137	TFDP2	rs347685	3.00E-11	67,093	22,982
Kottgen A	4q21.1	4	77368847	SHROOM3, CCDC158	rs17319721	1.00E-19	67,093	22,982
Kottgen A	5p13.1	5	39397132	DAB2,C9	rs11959928	1.00E-07	67,093	22,982
Kottgen A	5q35.3	5	176817636	SLC34A1,GRK6,RGS14,LMAN2,PRR7,F12,PFN3	rs6420094	1.00E-14	67,093	22,982
Kottgen A	6p21.1	6	43806609	VEGFA	rs881858	9.00E-14	67,093	22,982
Kottgen A	7q36.1	7	151407801	PRKAG2	rs7805747	4.00E-12	67,093	22,982
Kottgen A	8p21.2	8	23751151	STC1	rs10109414	1.00E-08	67,093	22,982
Kottgen A	9q21.11	9	71434707	PIP5K1B,FAM122A	rs4744712	8.00E-14	67,093	22,982
Kottgen A	12q24.12	12	112007756	ATXN2	rs653178	4.00E-11	67,093	22,982
Kottgen A	13q21.33	13	72347696	DACH1	rs626277	3.00E-11	67,093	22,982
Kottgen A	15q24.2	15	76158983	UBE2Q2,FBX022	rs1394125	3.00E-17	67,093	22,982
Kottgen A	16p12.3	16	20367690	UMOD,ACSM5,GP2,PDILT	rs12917707	1.00E-20	67,093	22,982
Kottgen A	19q13.11	19	33356891	SLC7A9,CCDC123,ECAT8	rs12460876	3.00E-15	67,093	22,982

Table 27: Demographic characteristics of the study population (Go-DARTS data genotyped on Affymetrix platform)

Study Characteristic	Mean(±Stdev)
Age at baseline	59.1(±11)
Sex (Female %)	46.4%
Baseline BMI	30.6(±5.3)
Baseline eGFR	73.9(±18.7)
Baseline Systolic BP	142.8(±18.4)
Baseline HbA1C	7.54(±1.3)
Baseline cholesterol	4.40(±0.97)
Duration of Diabetes at the end of follow-up	13.76(±7.74)
Duration of Follow-up(IQR)	10.6(±9.1)

Table 28: Baseline characteristics of cases with progression to CKD Stage 3B CKD and Controls

	CKD Stage 3B Case	Control	P-value
Age	60.45(10.0)	58.7(11.28)	0.001
Sex % Females	46.75%	47.63%	0.72
BMI	30.8(6.6)	29.2(8.7)	0.0001
Cholesterol	1.72(5.3)	2.76(4.5)	0.0001
HbA1c	5.4(5.5)	4.6(6.6)	0.005
Systoloic BP	123.7(52.02)	114.9(63.3)	0.001
Duration of Diabetes	15.9(7.9)	13.1(7.5)	0.0001
First eGFR	77.6(15.0)	72.8(18.8)	0.0001

Table 29: Association of the known SNPs with baseline eGFR, eGFR stratified by albuminuria status and time to stage 3B CKD

				Association with baseline eGFR(n=2970)		Association with eGFR in patients with sustained normoalbuminuria**(n=613)		Association with eGFR in patients with albuminuria(n=2097)		Interaction term Heterogeneity P-value	Association with time to Stage 3B CKD(eGFR<45)¥*		Direction of effect in Go-DARTS consistent with Anna Kottgen et al
CHR	Gene	SNP	Effect Allele	Beta(SE)	P-value	Beta(SE)	P-value	Beta(SE)	P-value		HR (CL)	P-value	
1	LASS2	rs267734	C	0.77(±0.51)	1.30E-01	2.24(±1.07)	3.63E-02	0.71(±0.62)	2.57E-01	9.60E-02	1.12(0.98,1.29)	7.00E-02	Yes
2	**GCKR**	**rs1260326**	**T**	**1.30(±0.44)**	**3.23E-03**	**0.45(±0.89)**	**6.12E-01**	**1.12(±0.55)**	**4.27E-02**	**8.70E-02**	**0.98(0.86,1.11)**	**7.60E-01**	**Yes**
2	NAT8	rs13538	G	0.40(±0.51)	4.32E-01	0.55(±1.12)	6.24E-01	0.29(±0.62)	6.34E-01	8.92E-01	1.02(1.023,1.027)	2.70E-01	Yes
3	TFDP2	rs347685	C	-0.51(±0.48)	2.82E-01	0.54(±0.97)	5.77E-01	-1.07(±0.59)	6.76E-02	3.95E-01	0.96(0.83,1.10)	5.50E-01	No
4	**SHROOM3**	**rs17319721**	**A**	**-1.28(±0.43)**	**3.18E-03**	**-0.07(±0.89)**	**9.34E-01**	**-1.43(±0.53)**	**7.28E-03**	**3.00E-03**	**1.02(0.90,1.15)**	**6.90E-01**	**Yes**
5	DAB2	rs11959928	A	-0.43(±0.45)	3.39E-01	-1.45(±0.90)	1.07E-01	-0.29(±0.55)	5.99E-01	3.41E-01	0.97(0.86,1.10)	7.00E-01	Yes
5	SLC34A1	rs6420094	G	-1.35(±0.61)	2.74E-02	-2.92(±1.24)	1.87E-02	-0.69(±0.75)	3.60E-01	2.79E-01	0.93(0.78.1.10)	4.00E-01	Yes
6	VEGFA	rs881858	G	0.54(±0.48)	2.63E-01	1.31(±1.01)	1.92E-01	1.34(±0.59)	2.21E-02	4.40E-02	0.95(0.83,1.08)	4.70E-01	Yes
7	PRKAG2	rs7805747	A	-0.31(±0.49)	5.24E-01	-0.72(±0.98)	4.62E-01	0.31(±0.60)	6.02E-01	9.30E-01	1.03(0.90.1.19)	6.00E-01	Yes
8	ADAM28	rs10109414	T	-0.51(±0.44)	2.41E-01	-1.57(±0.90)	8.17E-02	-0.17(±0.54)	7.49E-01	5.10E-01	0.99(0.87,1,12)	8.70E-01	Yes
9	PIP5K1B	rs4744712	A	0.09(±0.44)	8.47E-01	1.71(±0.91)	6.25E-02	-0.33(±0.55)	5.41E-01	9.31E-01	0.85(0.75,0.96)	1.00E-02	No
12	ATXN2	rs653178	T	0.20(±0.42)	6.28E-01	0.71(±0.85)	4.05E-01	-0.13(±0.52)	8.09E-01	9.47E-01	0.95(0.83,1.08)	9.50E-01	Yes
13	DACH1	rs626277	C	0.75(±0.44)	9.14E-02	0.85(±0.90)	3.46E-01	0.28(±0.54)	6.02E-01	3.93E-01	0.98(0.87,1.10)	7.50E-01	Yes
15	UBE2Q2	rs1394125	A	-0.86(±0.53)	1.03E-01	-1.14(±1.07)	2.89E-01	-0.86(±0.65)	1.85E-01	2.68E-01	1.11(0.96,1.28)	1.50E-01	Yes
16	**UMOD**	**rs12917707**	**T**	**2.0(±0.60)**	**8.84E-04**	**4.03(±1.23)**	**1.10E-03**	**1.72(±0.76)**	**2.30E-02**	**2.00E-03**	**0.83(0.70.0.99)**	**3.00E-02**	**Yes**
19	SLC7A9	rs12460876	C	0.24(±0.51)	6.90E-01	0.58(±0.94)	5.30E-01	0.29(±0.57)	6.00E-01	4.50E-01	0.86(0.76,0.98)	2.00E-02	Yes

**Adjusted for age at baseline, durdiab, baseline-eGFR, systolicbp, averagedHbA1c, averagedBMI*

***Patients with normo-albuminuria at baseline and at the end of follow-up with a duration of diabetes>15 years* ¥* Stage 3B CKD defined as 3 consecutive readings of eGFR <45. Those already at stage 3B CKD at baseline were excluded for this analysis

Table 30: Association of eGFR loci with Albuminuria in the Go-DARTS Cohort

CHR	SNP	BP	A1	BETA	SE	L95	U95	P	Gene
1	rs267734	149218101	C	0.705	0.6214	-0.5128	1.923	0.2567	LASS2
2	rs1260326	27584444	T	1.12	0.5525	0.03759	2.203	0.04269	GCKR
2	rs13538	73721836	G	0.2939	0.6173	-0.916	1.504	0.6341	NAT8
3	rs347685	143289827	C	-1.073	0.5865	-2.222	0.0769	0.06757	TFDP2
4	rs17319721	77587871	A	-1.431	0.5328	-2.476	-0.3871	0.007277	SHROOM3
5	rs11959928	39432889	A	-0.2898	0.5515	-1.371	0.7912	0.5993	DAB2
5	rs6420094	176750242	G	-0.6862	0.7496	-2.155	0.783	0.3602	SLC34A1
6	rs881858	43914587	G	1.344	0.5868	0.1941	2.494	0.02208	VEGFA
7	rs7805747	151038734	A	0.3142	0.6027	-0.867	1.496	0.6022	PRKAG2
8	rs10109414	23807096	T	-0.1728	0.5398	-1.231	0.8851	0.7489	ADAM28
9	rs4744712	70624527	A	-0.3342	0.5459	-1.404	0.7357	0.5405	PIP5K1B
12	rs653178	110492139	T	-0.1263	0.5217	-1.149	0.8962	0.8087	ATXN2
13	rs626277	71245697	C	0.2816	0.5397	-0.7762	1.339	0.6018	DACH1
15	rs1394125	73946038	A	-0.8559	0.646	-2.122	0.4102	0.1854	UBE2Q2
16	rs12917707	20275191	T	1.723	0.7574	0.2388	3.208	0.023	UMOD
19	rs12460876	38048731	C	0.2992	0.5759	-0.8295	1.428	0.6034	SLC7A9

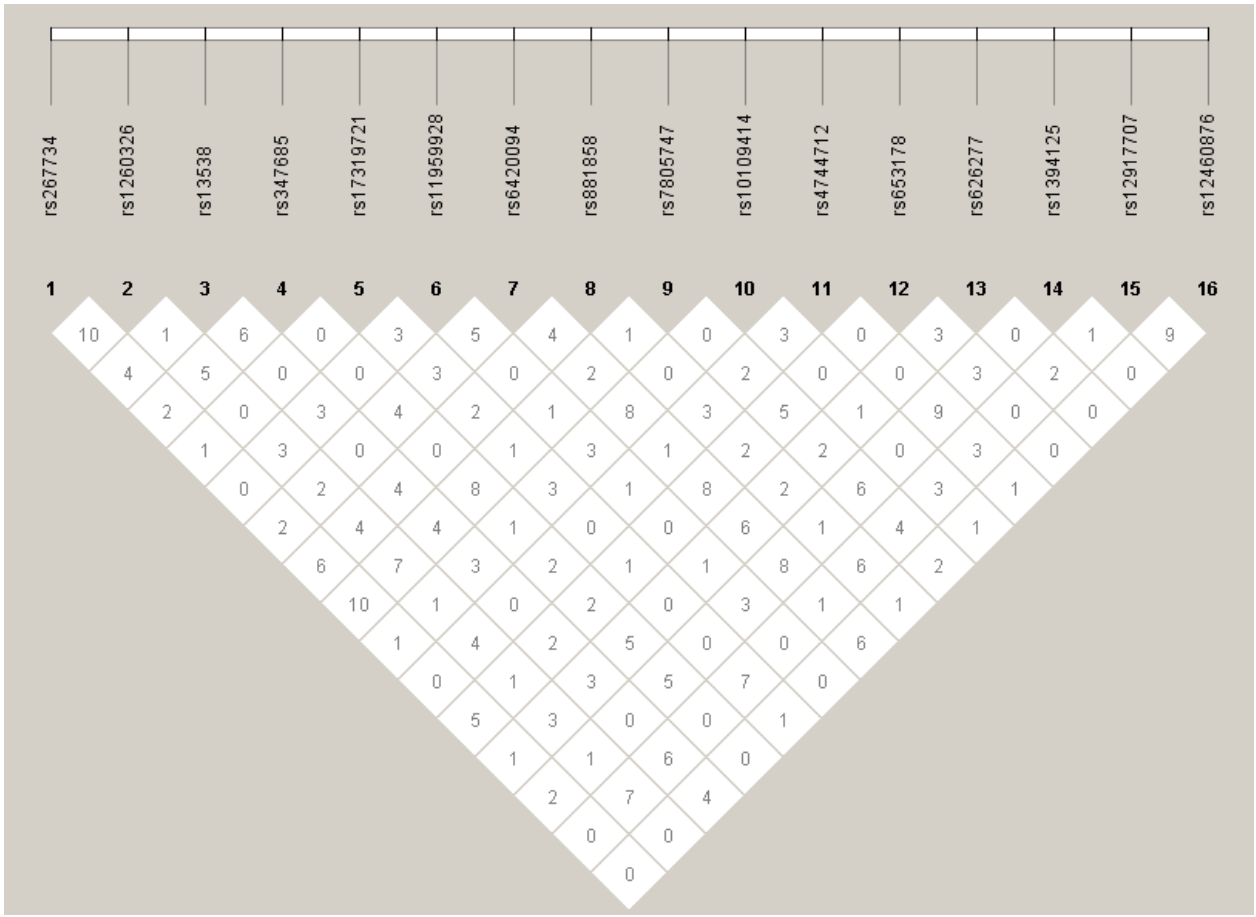


Figure 12 : Linkage Disequilibrium plot of the 16 known SNPs included in the study looking for association with eGFR

Figures in the squares indicate r-squared between the SNPs. All the r-squared<20 indicating that the SNPS are not in short or long range LD with each other.

Results II: Contribution to CKDgen consortium II

The contribution to the second round of analysis required estimation of eGFR from the MDRD equation and analysing eGFR as a continuous variable and as dichotomous variable (CKD60=eGFR<60 vs eGFR>60 and CKD45=eGFR <45 vs eGFR >45). Age and sex adjusted residuals for eGFR were used as quantitative trait (eGFRcrea). Results were stratified by age, sex and hypertension status. eGFR data was available for 2893 individuals while the CKD60 phenotype had 427 cases and 2466 controls case control and the CKD45 phenotype had N= 195 cases and 2466 controls. Here overall results for the three phenotypes are shown. Table 27 shows association of 35 SNPs with the CKD-60 phenotype, Table 28 shows the association of the 35 SNPs with CKD-45 phenotype and Table 29 shows the association of the 35 SNPs with eGFR as a quantitative trait.

Table 31: Replication for 35 SNPS for CKD 60 phenotype (N= 427 cases and 2466 controls) in Go-DARTS cohort

CHR	SNP	Pos	coded_allele	beta	se	pval	allele_freq_coded _allele	oevar_imp	Nearby_Gene
1	rs12124078	15742486	G	-0.12	0.09	0.15	0.31	1.00	*DNAJC16*
2	rs6431731	15780453	C	-0.14	0.20	0.49	0.05	0.86	*NA*
2	rs13014379	28966356	C	0.02	0.11	0.86	0.18	0.97	*NA*
2	rs4549145	48555686	G	-0.06	0.10	0.59	0.20	0.97	*CCDC128*
2	rs10490130	169807357	C	0.06	0.15	0.69	0.07	0.99	*LRP2*
2	rs6433115	169899211	C	-0.04	0.10	0.65	0.21	1.00	*LRP2*
2	rs11683577	170004477	T	-0.03	0.09	0.71	0.31	0.99	*NA*
3	rs9827843	76557775	G	0.03	0.08	0.74	0.39	1.00	*NA*
3	rs9824190	98163910	G	0.21	0.15	0.18	0.06	1.00	*EPHA6*
4	rs11935537	130709590	G	0.11	0.11	0.32	0.15	0.99	*NA*
5	rs10068737	118064446	C	-0.03	0.08	0.69	0.48	1.00	*NA*
6	rs1264701	30174337	T	-0.21	0.12	0.07	0.15	0.96	*NA*
6	rs1322199	165345507	G	-0.04	0.09	0.62	0.29	1.00	*NA*
7	rs500456	54566276	G	-0.07	0.09	0.40	0.40	0.98	*NA*
8	rs1500896	96555858	A	-0.13	0.08	0.11	0.42	0.99	*NA*
9	rs2184241	3574112	T	0.02	0.10	0.80	0.36	0.94	*NA*
9	rs4149333	106592745	G	0.03	0.11	0.76	0.13	0.99	*ABCA1*
9	rs1050700	134757764	C	0.16	0.09	0.07	0.28	0.97	*TSC1*
10	rs7911360	122795537	A	-0.09	0.08	0.24	0.39	1.00	*NA*
10	rs4751890	124151781	C	0.06	0.08	0.48	0.42	1.00	*PLEKHA1*
10	rs11245299	126253483	A	0.04	0.09	0.68	0.27	1.00	*LHPP*
11	rs6421967	973966	T	0.01	0.08	0.90	0.36	0.99	*AP2A2*
11	rs3925584	30716911	C	-0.19	0.08	0.02	0.46	1.00	*NA*
11	rs489381	62408639	A	-0.11	0.15	0.48	0.08	0.95	*SLC3A2 S*
11	rs752805	125798495	A	0.07	0.08	0.40	0.48	1.00	*KIRREL3*
12	rs3741414	56130316	T	0.04	0.09	0.63	0.23	0.99	*INHBC*
15	rs2928148	39188842	G	0.07	0.08	0.35	0.46	0.99	*INOC1*
15	rs9302109	39399472	T	0.01	0.08	0.92	0.33	1.00	*OIP5*
15	rs7176121	98662843	T	-0.09	0.08	0.26	0.47	1.00	*ADAMTS17*
16	rs249942	23552849	A	-0.14	0.20	0.50	0.05	0.93	*PALB2*
16	rs6499166	66884418	G	0.02	0.09	0.80	0.24	0.99	*SLC7A6*
17	rs2453580	19378913	C	-0.12	0.13	0.33	0.31	0.92	*SLC47A1*
17	rs7208487	34796975	G	0.01	0.11	0.91	0.15	0.99	*FBXL20*
17	rs11078903	34885450	G	-0.11	0.12	0.36	0.17	0.97	*CRKRS*
18	rs17446008	23875364	A	0.25	0.32	0.43	0.02	0.70	*CDH2*

Table 32: Replication for 35 SNPS for CKD 45 phenotype (N= 195 cases and 2466 controls) in Go-DARTS cohort

CHR	SNP	Pos	coded_allele	beta	se	pval	allele_freq_coded _allele	oevar_imp	*Nearby_Gene*
1	rs12124078	15742486	G	0.10	0.16	0.53	0.32	1.00	*DNAJC16*
2	rs6431731	15780453	C	-0.26	0.43	0.55	0.04	0.86	*NA*
2	rs13014379	28966356	C	0.11	0.21	0.60	0.18	0.97	*NA*
2	rs4549145	48555686	G	0.06	0.20	0.78	0.20	0.97	*CCDC128*
2	rs10490130	1.7E+08	C	-0.03	0.29	0.91	0.07	0.99	*LRP2*
2	rs6433115	1.7E+08	C	-0.40	0.20	0.05	0.21	1.00	*LRP2*
2	rs11683577	1.7E+08	T	0.00	0.17	0.99	0.31	0.99	*NA*
3	rs9827843	76557775	G	-0.23	0.16	0.13	0.40	1.00	*NA*
3	rs9824190	98163910	G	0.37	0.27	0.17	0.07	1.00	*EPHA6*
4	rs11935537	1.31E+08	G	0.26	0.23	0.26	0.15	0.99	*NA*
5	rs10068737	1.18E+08	C	-0.10	0.16	0.54	0.46	1.00	*NA*
6	rs1264701	30174337	T	-0.29	0.23	0.21	0.15	0.96	*NA*
6	rs1322199	1.65E+08	G	0.06	0.16	0.71	0.30	1.00	*NA*
7	rs500456	54566276	G	-0.28	0.18	0.12	0.40	0.98	*NA*
8	rs1500896	96555858	A	-0.04	0.16	0.77	0.43	0.99	*NA*
9	rs2184241	3574112	T	0.02	0.21	0.94	0.36	0.94	*NA*
9	rs4149333	1.07E+08	G	0.26	0.22	0.25	0.12	0.99	*ABCA1*
9	rs1050700	1.35E+08	C	0.00	0.17	0.98	0.31	0.97	*TSC1*
10	rs7911360	1.23E+08	A	0.15	0.16	0.34	0.40	1.00	*NA*
10	rs4751890	1.24E+08	C	0.11	0.16	0.47	0.42	1.00	*PLEKHA1*
10	rs11245299	1.26E+08	A	-0.02	0.18	0.93	0.26	1.00	*LHPP*
11	rs6421967	973966	T	0.20	0.15	0.19	0.37	0.99	*AP2A2*
11	rs3925584	30716911	C	-0.30	0.16	0.06	0.46	1.00	*NA*
11	rs489381	62408639	A	-0.10	0.29	0.72	0.08	0.95	*SLC3A2 S*
11	rs752805	1.26E+08	A	0.09	0.16	0.56	0.47	1.00	*KIRREL3*
12	rs3741414	56130316	T	-0.21	0.19	0.27	0.23	0.99	*INHBC*
15	rs2928148	39188842	G	0.18	0.15	0.23	0.45	0.99	*INOC1*
15	rs9302109	39399472	T	-0.13	0.17	0.45	0.32	1.00	*OIP5*
15	rs7176121	98662843	T	0.32	0.15	0.03	0.48	1.00	*ADAMTS17*
16	rs249942	23552849	A	0.75	0.34	0.03	0.05	0.93	*PALB2*
16	rs6499166	66884418	G	-0.30	0.19	0.11	0.24	0.99	*SLC7A6*
17	rs2453580	19378913	C	0.00	0.26	0.99	0.30	0.92	*SLC47A1*
17	rs7208487	34796975	G	0.19	0.21	0.38	0.16	0.99	*FBXL20*
17	rs11078903	34885450	G	0.52	0.21	0.01	0.17	0.97	*CRKRS*
18	rs17446008	23875364	A	-0.31	0.80	0.70	0.02	0.70	*CDH2*

Table 33: Replication for 35 SNPS for eGFR phenotype (N=2893) in Go-DARTS cohort

CHR	SNP	Pos	coded_allele	strand	beta	se	pval	allele_freq_coded _allele	oevar_imp	*Nearby_Gene*
1	rs12124078	15742486	G	+	0.00	0.01	8.19E-01	0.31	1.00	*DNAJC16*
2	rs6431731	15780453	C	+	0.01	0.02	4.87E-01	0.05	0.86	*NA*
2	rs13014379	28966356	C	+	0.00	0.01	9.97E-01	0.18	0.97	*NA*
2	rs4549145	48555686	G	+	0.01	0.01	2.71E-01	0.20	0.97	*CCDC128*
2	rs10490130	169807357	C	+	-0.01	0.02	5.22E-01	0.07	0.99	*LRP2*
2	rs6433115	169899211	C	+	0.01	0.01	1.50E-01	0.21	1.00	*LRP2*
2	rs11683577	170004477	T	+	0.01	0.01	2.39E-01	0.31	0.99	*NA*
3	rs9827843	76557775	G	+	0.00	0.01	7.56E-01	0.39	1.00	*NA*
3	rs9824190	98163910	G	+	-0.02	0.02	1.76E-01	0.06	1.00	*EPHA6*
4	rs11935537	130709590	G	+	0.00	0.01	8.33E-01	0.15	0.99	*NA*
5	rs10068737	118064446	C	+	0.00	0.01	9.56E-01	0.48	1.00	*NA*
6	rs1264701	30174337	T	+	0.03	0.01	1.72E-02	0.15	0.96	*NA*
6	rs1322199	165345507	G	+	0.00	0.01	9.12E-01	0.29	1.00	*NA*
7	rs500456	54566276	G	+	0.01	0.01	4.58E-01	0.40	0.98	*NA*
8	rs1500896	96555858	A	+	0.01	0.01	5.16E-01	0.42	0.99	*NA*
9	rs2184241	3574112	T	+	0.01	0.01	4.43E-01	0.36	0.94	*NA*
9	rs4149333	106592745	G	+	0.00	0.01	9.10E-01	0.13	0.99	*ABCA1*
9	rs1050700	134757764	C	+	-0.01	0.01	2.89E-01	0.28	0.97	*TSC1*
10	rs7911360	122795537	A	+	0.01	0.01	1.99E-01	0.39	1.00	*NA*
10	rs4751890	124151781	C	+	0.00	0.01	6.75E-01	0.42	1.00	*PLEKHA1*
10	rs11245299	126253483	A	+	-0.01	0.01	4.98E-01	0.27	1.00	*LHPP*
11	rs6421967	973966	T	+	0.00	0.01	6.54E-01	0.36	0.99	*AP2A2*
11	rs3925584	30716911	C	+	0.03	0.01	8.24E-04	0.46	1.00	*NA*
11	rs489381	62408639	A	+	-0.01	0.02	4.20E-01	0.08	0.95	*SLC3A2 S*
11	rs752805	125798495	A	+	0.00	0.01	6.71E-01	0.48	1.00	*KIRREL3*
12	rs3741414	56130316	T	+	0.01	0.01	3.97E-01	0.23	0.99	*INHBC*
15	rs2928148	39188842	G	+	-0.01	0.01	3.40E-01	0.46	0.99	*INOC1*
15	rs9302109	39399472	T	+	0.00	0.01	6.78E-01	0.33	1.00	*OIP5*
15	rs7176121	98662843	T	+	0.00	0.01	8.97E-01	0.47	1.00	*ADAMTS17*
16	rs249942	23552849	A	+	0.01	0.02	7.80E-01	0.05	0.93	*PALB2*
16	rs6499166	66884418	G	+	-0.01	0.01	5.07E-01	0.24	0.99	*SLC7A6*
17	rs2453580	19378913	C	+	0.00	0.01	9.25E-01	0.31	0.92	*SLC47A1*
17	rs7208487	34796975	G	+	0.00	0.01	8.27E-01	0.15	0.99	*FBXL20*
17	rs11078903	34885450	G	+	-0.01	0.01	5.69E-01	0.17	0.97	*CRKRS*
18	rs17446008	23875364	A	+	-0.02	0.04	5.87E-01	0.02	0.69	*CDH2*

V. GWAS for DKD in EURODIAB and GO-DARTS dataset

In the following section, we show summary of the results of the GWAS analysis for DKD phenotypes in EURODIAB and Go-DARTS data set. The intention of the analysis was to contribute data to the SUMMIT GWAS meta-analysis. The results below show q-q plot of the five DKD phenotypes and the top hit in each of the phenotypes. Most of the top hits show MAF of <5% and are likely to be false positive, given the low power of this single study to detect rare variants. We have not discussed these results in details as these analyses were primarily done to perform meta-analysis with other SUMMIT consortium datasets. Table 30 shows sample sizes for T2D DKD Phenotypes in the GO-DARTS Affymetrix and Go-DARTS Illumina cohorts.

The sample sizes (Case/Control) for Go-DARTS DKD GWAS were 1744/1496 for Any Albuminuria, 1997/2066 for CKD, 1347/1496 for Micro-albuminuria, 397/1496 for Macro-albuminuria , 128/1496 for ESRD and 288/1303 for Macroalbuminuria&CKD phenotypes.

The Sample sizes for (Case/Control) for Eurodiab DKD GWAS were 298/491for Any Albuminuria, 113/467 for CKD, 95/491 for Micro-albuminuria, 203/491 for Macro-albuminuria, 84/491 for ESRD and 210/357 for Macroalbuminuria&CKD phenotypes. Analysis was performed in SNPTEST for all the GWAS studies adjusting for age, sex, and duration of diabetes. The SNPTEST output was then analysed with R to perform quality control before the meta-analysis. The R script for quality control is listed in Appendix A; it briefly consisted of following steps

a) Calculate the effect allele frequency

b) Remove SNP with MAF<1%

c) Remove SNP with HWE P-value<10-5

d) Remove SNP with information content <0.40

Once the quality control was performed Manhattan plots and Q-Q plots were drawn using R scripts listed in Appendix 1.

Figure 13: QQ-Plots of DN phenotype in EURODIAB dataset

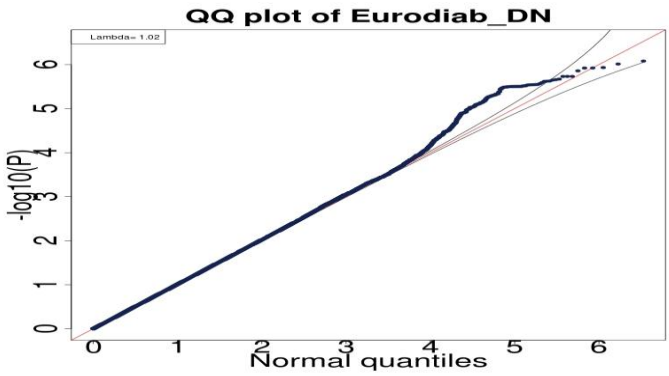


Figure 14: QQ-Plots of CKD phenotype in EURODIAB dataset

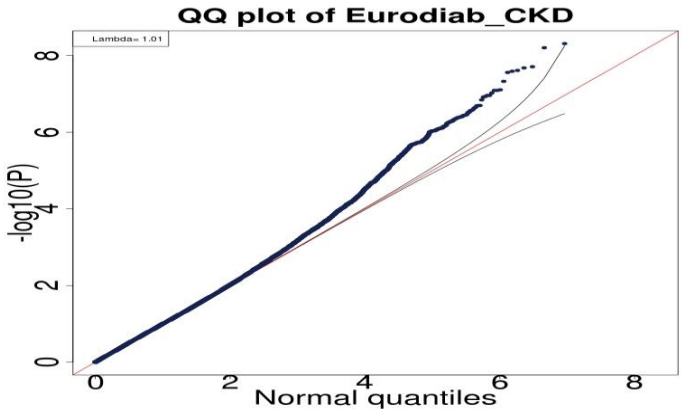


Figure 15: QQ-Plots of CKD-DN phenotype in EURODIAB dataset

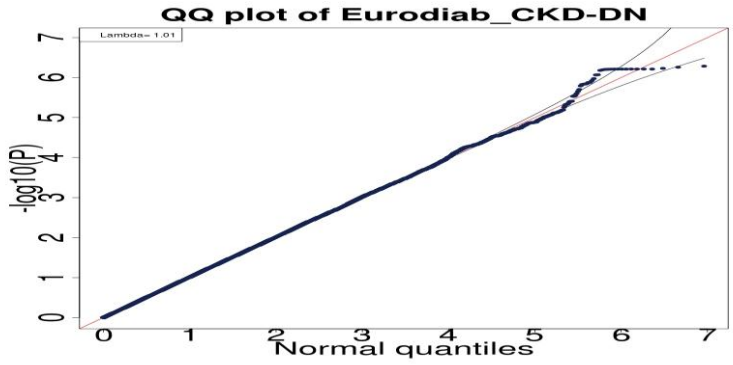


Figure 16: QQ-Plots of Macro-albuminuria phenotype in EURODIAB dataset

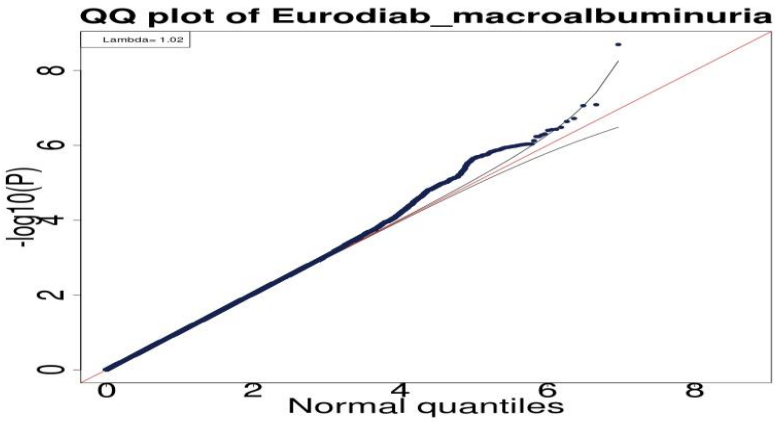


Figure 17: QQ-Plots of Micro-albuminuria phenotype in EURODIAB dataset

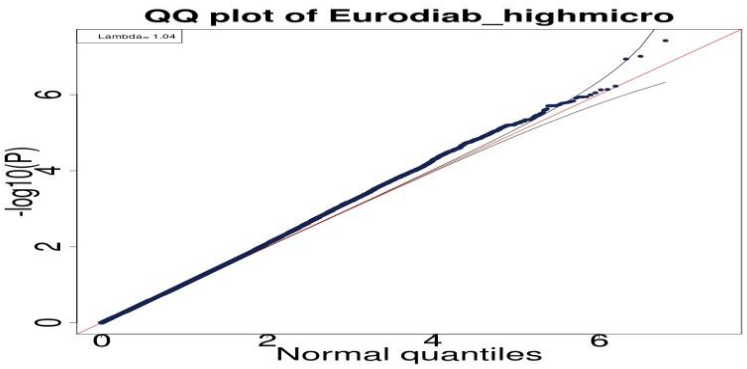


Figure 18: QQ-Plots of ESRD phenotype in EURODIAB dataset

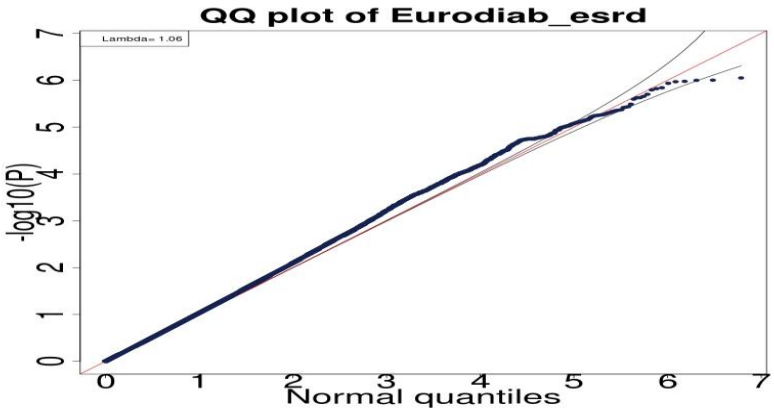


Figure 19: QQ-Plots of DN phenotype in GO-DARTS dataset

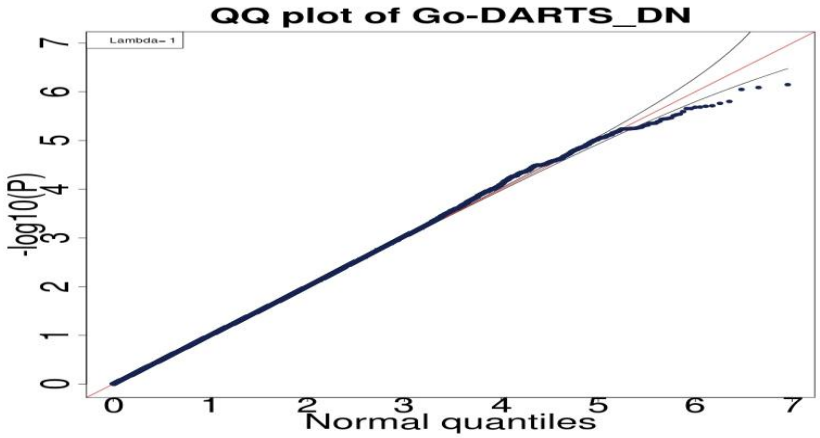


Figure 20: QQ-Plots of CKD phenotype in GO-DARTS dataset

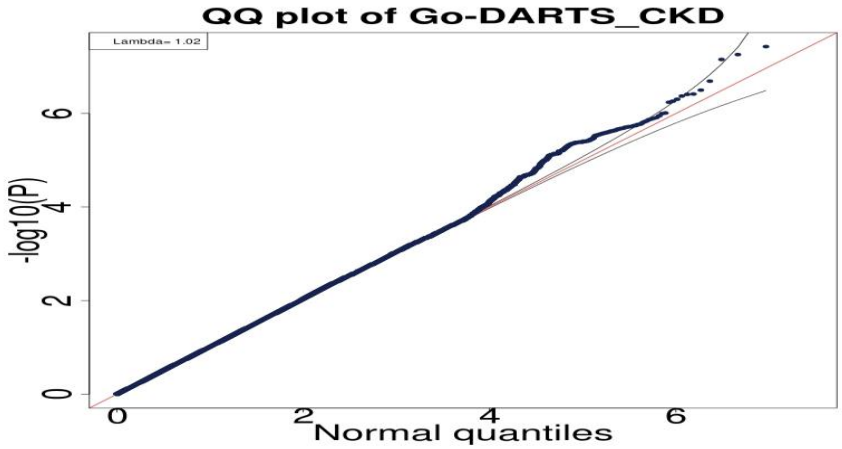


Figure 21: QQ-Plots of CKD-DN phenotype in GO-DARTS dataset

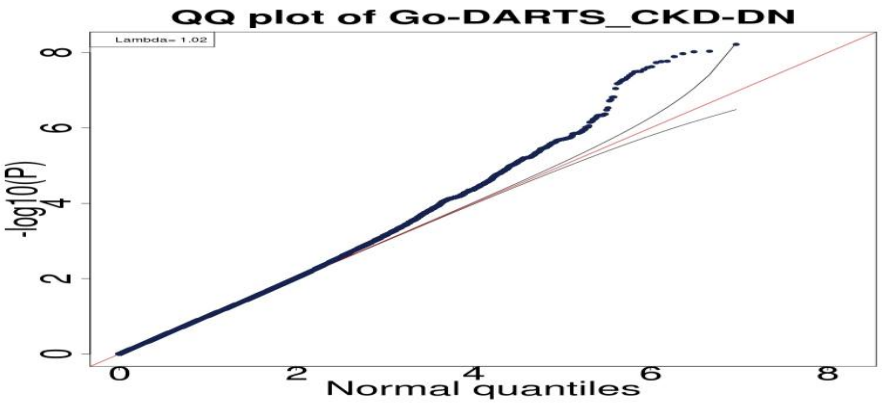


Figure 22: QQ-Plots of Macro-albuminuria phenotype in GO-DARTS dataset

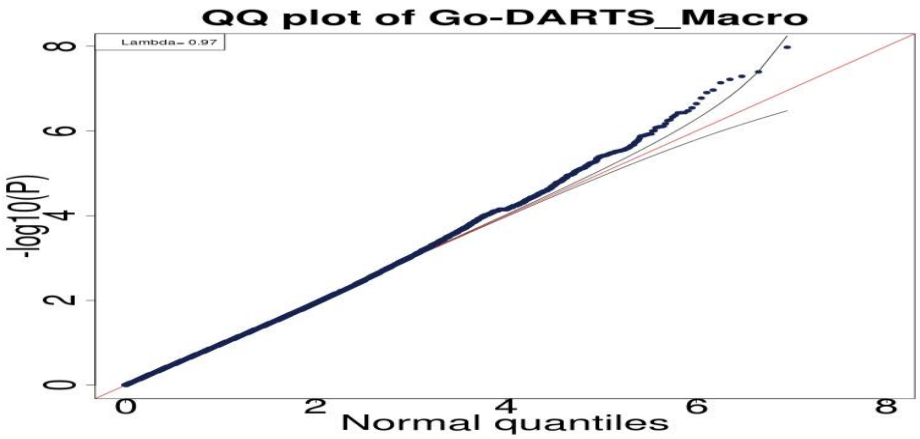


Figure 23: QQ-Plots of Micro-albuminuria phenotype in GO-DARTS dataset

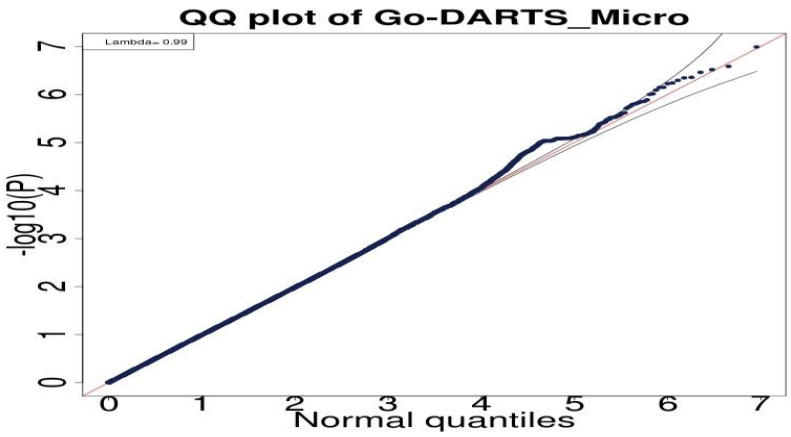


Figure 24: QQ-Plots of ESRD phenotype in GO-DARTS dataset

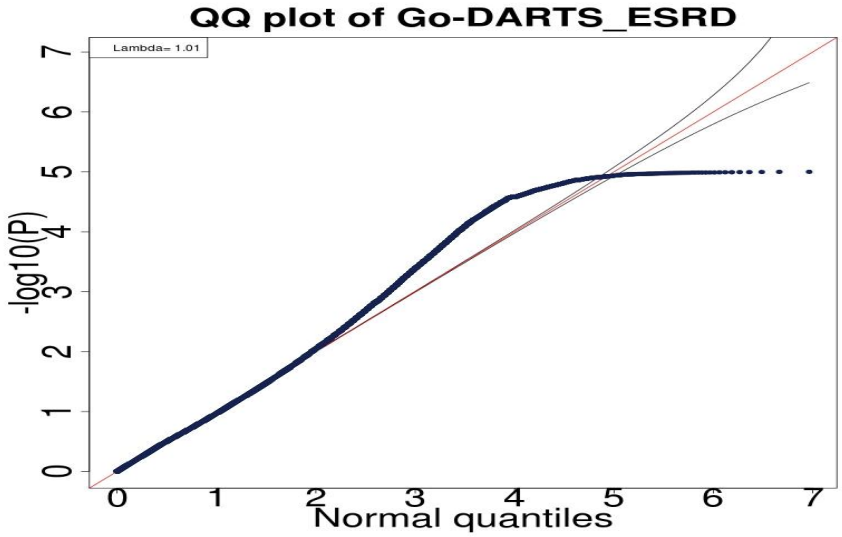


Table 34: Sample Size for T2D-DKD phenotypes

Phenotype	Diabetic Nephropathy		CKD		Micro-albuminuria		Macro-albuminuria		ESRD		Macroalbuminuria +CKD	
	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control
Godarts_affy	**885**	**816**	**1,025**	**1,553**	**667**	**816**	**218**	**816**	**80**	**816**	**168**	**716**
Steno	163	131	100	174	NA	NA	163	131	NA	NA	NA	NA
GoDARTS_illumina	**859**	**680**	**972**	**513**	**680**	**680**	**179**	**680**	**48**	**680**	**120**	**587**
MNI	188	165	NA	NA	122	162	66	165	NA	NA	NA	NA
SDR	1,250	580	997	666	520	580	713	580	243	580	609	307
Total	3,345	2,372	3,094	2,906	1,989	2,238	1339	2372	371	2,076	897	1,610

Table 35: Sample Size for T1D-DKD phenotypes

Phenotype	Diabetic Nephropathy		CKD		Micro-albuminuria		Macro-albuminuria		ESRD		Macroalbuminuria+CKD	
	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control
Eurodiab	**298**	**491**	**113**	**467**	**95**	**491**	**203**	**491**	**84**	**491**	**210**	**357**
SDR	266	290	163	365	98	290	168	292	75	294	118	239
FinnDiane	1,802	1,613	2,077	979	463	1,613	1,339	1,613	654	1,613	1,422	789
Cambridge	197	199	NA	NA	150	199	47	199	NA	NA	NA	NA
Total	2,563	2,593	2353	1811	806	2,593	1,757	2,595	813	2398	1,750	1,385

Table 36: Top GWAS hit for each of the DKD phenotypes in EURODIAB and Go-DARTS dataset

Study Cohort	Phenotype	CHR	Position	SNP	MAF	P-value	Nearby-Gene
EURODIAB	CKD	5	rs113471261	51701706	0.02	4.90E-09	*CTD-2288O8.1*
EURODIAB	CKD-DN	8	rs74888877	58650530	0.01	5.18E-07	*RP11-388G22.1*
EURODIAB	ESRD	15	rs2708291	67197785	0.04	8.91E-07	*RP11-798K3.4*
EURODIAB	Macro-ESRD	10	rs192513143	71060011	0.01	2.02E-09	*HK1*
EURODIAB	Micro-albuminuria	3	rs140827779	55464667	0.03	3.80E-08	*WNT5A*
EURODIAB	Any-albuminuria	4	rs6553996	31871630	0.28	8.30E-07	*RP11-734I18.1*
Go-DARTS	CKD	3	rs75224764	192126446	0.01	3.71E-08	*FGF12*
Go-DARTS	CKD-DN	2	rs181726291	102287994	0.01	6.05E-09	*MAP4KA*
Go-DARTS	ESRD	2	rs145414014	189576382	0.01	1.00E-04	*GULP1*
Go-DARTS	Macro-ESRD	10	rs41307583	127344489	0.03	1.06E-08	*HNF1*
Go-DARTS	Micro-albuminuria	19	chr19:2766244:D	2766244	0.01	1.03E-07	*NA*
Go-DARTS	Any-albuminuria	7	rs76562431	125478081	0.40	7.13E-07	*AC005276.1*

*Since the intention was to Meta-analyse the data with cohorts in SUMMIT DKD phenotypes only the top SNP for each phenotype with its P-value are shown

** All the SNPS had information content in SNPTEST>0.40

VI. Meta-analysis of GWAS data with other datasets

Tables 32 and 33 show baseline demographic characteristics of the GWAS studies for T1D-DKD and T2D-DKD used in the meta-analysis. Tables 34-45 and Figures 26-37 show Q-Q plot, Manhattan plot and top 5 signals from the GWAS meta-analysis of 6 DKD sub-phenotypes; CKD, CKD-DN, Macro-albumiuria+ESRD, ESRD, Micro-albuminuria and DN(any-albuminuria).

Top Results in T1D-DKD

The strongest association (P=5.96x10^{-7}) was observed for rs12632850 on chromosome 3q25.33 between *SCHIP1* (schwannomin interacting protein 1) and *IL12A* (interleukin 12A) genes. For the CKD-DN phenotype strong association was also seen for rs17023084 (P=4.08x10^{-8}), located in the *AFF3* gene. For the ESRD phenotype strongest association was also seen for rs76729345 in *INSR* (Insulin Receptor, P=4.40x10^{-6}) gene. For macro-albuminuria+ESRD phenotype a strong association (P=2.18x10^{-6}) was observed for rs681586 on chromosome 2 between *EN1* (engrailed homeobox 1) and *INSIG2* (insulin induced gene 2). For micro-albuminuria the strongest association (P=2.15x10^{-7}) was observed for rs62404695 in *LINC00340* (long intergenic non-protein coding RNA 340). For the main DN phenotype, strongest association was seen for rs11123857 on chromosome 2, P=4.6×10^{-7}. rs11123857 is located intronic in *NPAS2* gene (neuronal PAS domain protein 2).

Top Results in T2D-DKD

The strongest association with CKD was seen for rs2206136 in *PLCB4* is associated with CKD at genome-wide significance ($p=2.1x10^{-8}$). The strongest association with the combined CKD + Albuminuria phenotype is seen for rs6997279 ($p=2.2x10^{-7}$) in the *SLC30A8* gene. For the ESRD phenotype, two loci were associated at $p<10^{-6}$: rs2475363, near *EIF5AL1* and rs13224128, near *CHN2.* A less significant but biologically relevant signal was detected near *KLF10,* rs4734659 ($p=1.0x10^{-6}$). For macro-albuminuria+ ESRD phenotype three SNPs are associated with $p<10^{-6}$: rs116354014 ($p=2.5x10^{-7}$), near *TBC1D5*; rs112765093 ($p=2.9x10^{-7}$), near *GTDC1*; and rs147523203 ($p=5.9x10^{-7}$), near *GALNTL6*. For macro-albuminuria+ ESRD a biologically relevant association rs7942230 ($p=2.25x10^{-6}$) was seen at intronic SNP in the *GRM5* gene encoding glutamate receptor 5. The strongest association for micro-albuminuria was observed for rs2150814 ($p=8.1x10^{-8}$) situated between *GABRR1* and *GABRR2*. An additional three loci were associated with micro-albuminuria at $p <10^{-7}$, including rs1143914 in the intron of *COL4A1*. Two loci were associated with the primary phenotype (DN) with $p<10^{-6}$: Rs183249293 ($p=2.4x10^{-7}$), a rare variant (MAF=0.016) located in an intergenic region near *SLITRK6,* and rs9942471 ($p=4.8x0^{-7}$), a common variant (MAF=0.36) located between *GABRR1* and *GABRR2*. From the top 20 associations chr4: 55939605:I ($p=2.1x10^{-6}$) situated downstream of the *KDR* gene may be an interesting candidate locus

Table 37: Baseline characteristics of Type 1 Diabetes Cohort

	EURODIAB (n=1016)	**FinnDiane (n= 3435)**	**Scania Type 1 Diabetes (n=513)**	**Cambridge Type 1 Diabetes Cohort (n=400)**
Age in years	42(±10.3)	38.5(±12.3)	40.5(±12.3)	37.5(±11.3)
Age of onset of Type 1 diabetes	18(±6.2)	19(±4.2)	18(±8.4)	19(±8.3)
Sex (%Males)	50.30%	48.20%	45.20%	47.20%
BMI	24.8(±3.4)	25.1(±2.5)	26.1(±3.5)	24.1(±3.5)
HbA1c	7.8(±1.7)	8.3(±1.4%)	8.1(±1.3%)	8.2(±13%)
Duration of Diabetes in years	24.3(±8.3)	23.4(±6.8)	22.4(±12.9)	21.4(±12.8)
Number of SNPs	~9 million	~9 million	~9 million	~9 million

Table 38: Baseline characteristics of Type 2 diabetes cohort

	Go-DARTS I (n=1701)	**Go-DARTS 2 (n=1539)**	**Steno diabetes centre (n=294)**	**Scania Type 2 diabetes register (n=1830)**
Age in years	59.1 (11.0)	66.2 (11.6)	51.7(11.2)	53.7(11.5)
Sex (%Males)	46.40%	42.30%	56%	55%
BMI	30.6 (5.3)	31.5 (6.1)	30.1(5.8)	30.1(5.8)
HbA1c	7.54(1.3)	7.3(1.4)	6.2(1.6)	7.2(1.6)
Duration of Diabetes in years	8.71(7.44)	7.75(6.61)	7.1(5.5)	15.1(5.5)
Number of SNPS	~ 9 million	~ 9 million	~ 9 million	~ 9 million

Table 39: Meta-Analysis of CKD phenotype in SUMMIT consortium in patients with Type 1 Diabetes

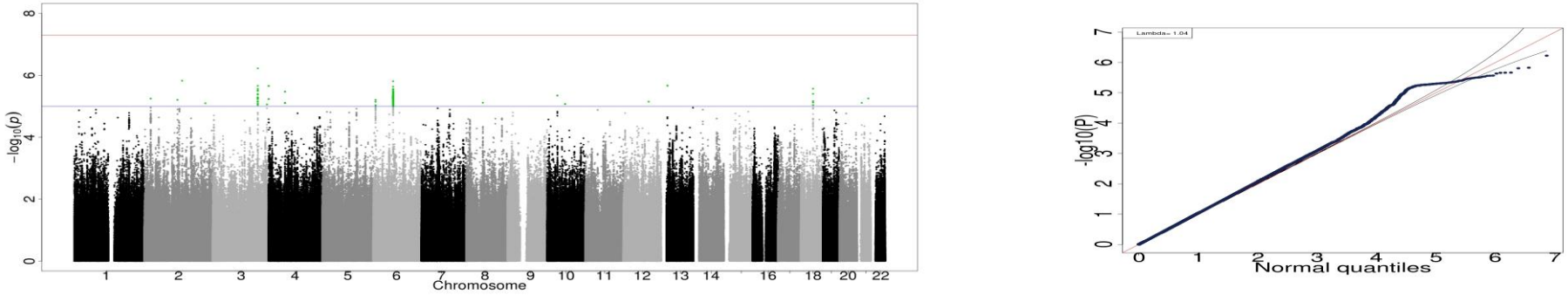


Figure 25: Manhattan Plot and QQ-plot SUMMIT CKD Phenotype T1D

RS_NUMBER	CHR	BP	P	eaf	reference_allele	OR_95U	OR	OR_95L	P	Nearest Gene
rs12632850	3	1.6E+08	5.96E-07	0.20	G	0.84	0.74	0.65	5.96E-07	*IQCJ-SCHIP1*
rs72842475	2	1.34E+08	1.49E-06	0.06	G	0.75	0.61	0.49	1.49E-06	*AC010890.1*
rs9342772	6	70461240	1.55E-06	0.38	C	0.86	0.78	0.70	1.55E-06	*NPM1P37*
rs9509964	13	22581435	2.16E-06	0.03	C	2.89	2.11	1.54	2.16E-06	*NME1P1*
rs13113236	4	424368	2.20E-06	0.17	T	0.83	0.72	0.62	2.20E-06	*AC092574.1*

Table 40: Meta-Analysis of CKD-DN phenotype in SUMMIT consortium in patients with Type 1 Diabetes

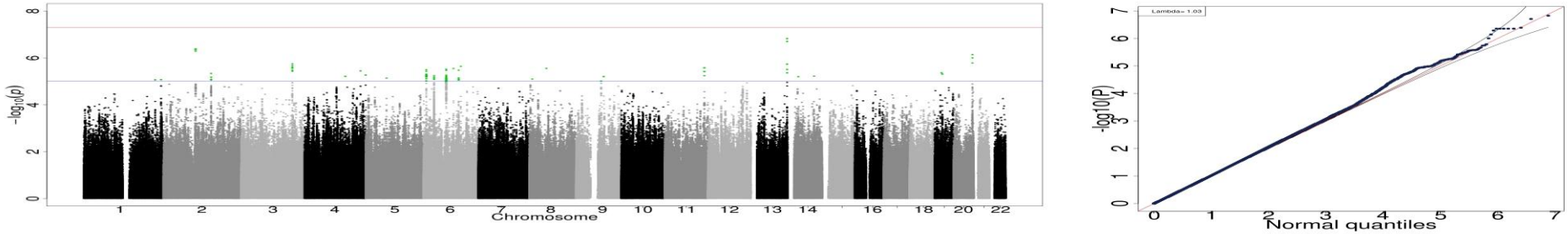


Figure 26: Manhattan Plot and Q-Q plot SUMMIT CKD-DN Phenotype Type 1 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	OR_95U	OR	OR_95L	Within Gene
rs75569002	13	1.13E+08	1.49E-07	0.13	T	1.89	1.59	1.33	*ATP11A*
rs17023084	2	1E+08	4.08E-07	0.05	G	2.59	2.00	1.54	*AFF3*
rs6027504	20	58893816	7.19E-07	0.43	C	0.83	0.74	0.66	*RP5-1043L13.1*
rs12632850	3	1.6E+08	1.79E-06	0.20	G	0.82	0.72	0.62	*NA*
rs77501404	6	1.16E+08	2.28E-06	0.05	T	2.40	1.85	1.43	*NA*

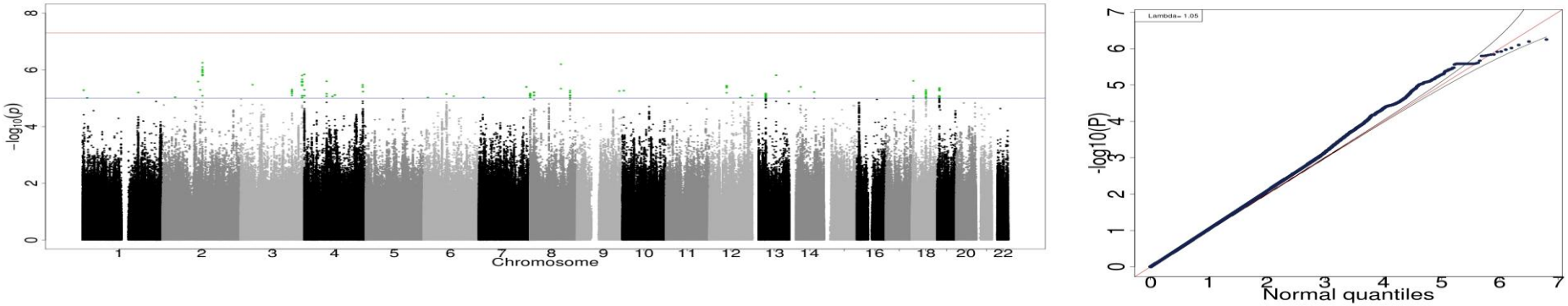


Figure 27: Manhattan Plot and Q-Q plot SUMMIT ESRD Phenotype Type 1 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	OR_95U	OR	OR_95L	Within Gene
rs72841506	2	1.24E+08	5.61E-07	0.06	A	2.39	1.86	1.45	*NA*
rs61675991	8	96399614	6.38E-07	0.10	T	2.12	1.72	1.39	*KB-1047C11.2*
rs13113236	4	424368	1.45E-06	0.18	T	0.80	0.67	0.56	*ZNF721*
rs58573474	13	73938534	1.55E-06	0.42	G	1.49	1.32	1.17	*NA*
rs78247479	3	1.91E+08	1.58E-06	0.06	C	2.32	1.80	1.39	*RP11-197K6.1*

Table 41: Meta-Analysis of ESRD phenotype in SUMMIT consortium in patients with Type 1 Diabetes

Table 42: Meta-Analysis of Macro-ESRD phenotype in SUMMIT consortium in patients with Type 1 Diabetes

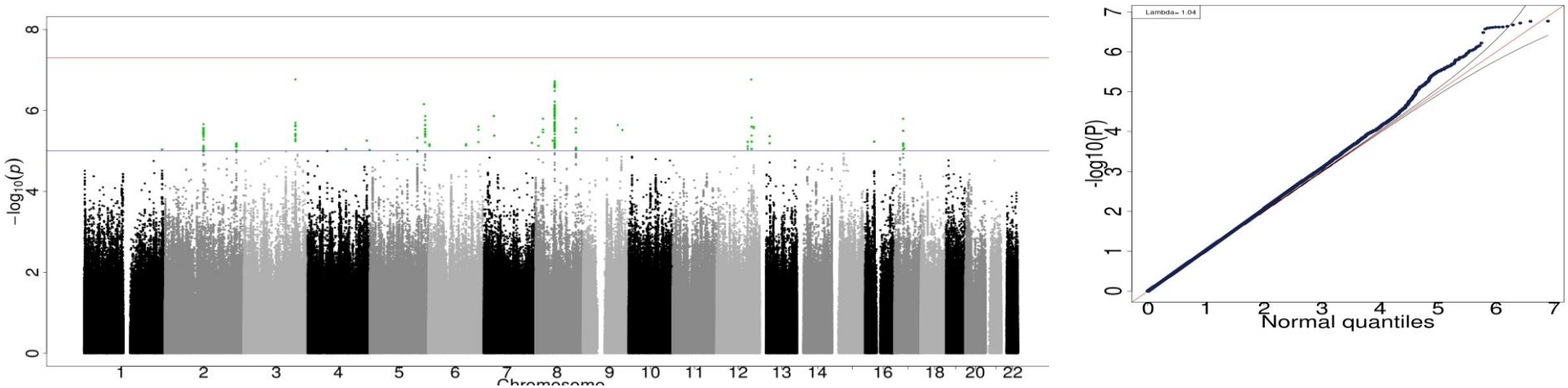


Figure 28: Manhattan Plot and Q-Q plot SUMMIT Macro-ESRD Phenotype Type 1 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	OR_95U	OR	OR_95L	Within Gene
rs12632850	3	1.6E+08	1.71E-07	0.20	G	0.82	0.73	0.65	*IQCJ-SCHIP1*
rs145462438	12	1.07E+08	1.73E-07	0.02	A	4.14	2.81	1.91	*C12orf23*
rs2970761	8	59279292	1.91E-07	0.77	G	1.48	1.33	1.19	*RP11-114M5.1*
rs880057	8	59135088	2.12E-07	0.81	A	1.51	1.34	1.19	*FAM110B*
rs72831309	5	1.67E+08	6.99E-07	0.04	A	2.64	2.00	1.52	*TENM2*

Table 43: Meta-Analysis of Micro phenotype in SUMMIT consortium in patients with Type 1 Diabetes

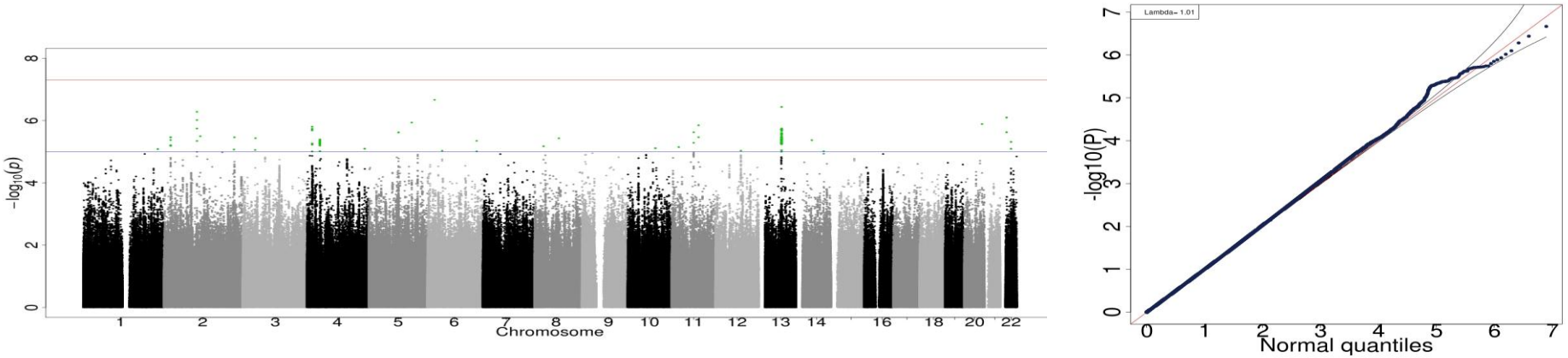


Figure 29: Manhattan Plot and Q-Q plot SUMMIT Micro-albuminuria Phenotype Type 1 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	OR_95U	OR	OR_95L	Within Gene
rs62404695	6	22100761	2.15E-07	0.02	C	7.04	4.15	2.45	*LINC00340*
chr13:70304398:I	13	70304398	3.64E-07	0.05	AAT	2.84	2.07	1.51	*KLHL1*
rs56003443	2	1.02E+08	5.22E-07	0.22	A	1.64	1.42	1.23	*NPAS2*
rs75270220	22	19989739	7.97E-07	0.02	C	10.49	5.53	2.91	*ARVCF*
rs60813019	5	1.32E+08	1.16E-06	0.05	A	2.95	2.15	1.57	*AC010240.3*

Table 44: Meta-Analysis of DN phenotype in SUMMIT consortium in patients with Type 1 Diabetes

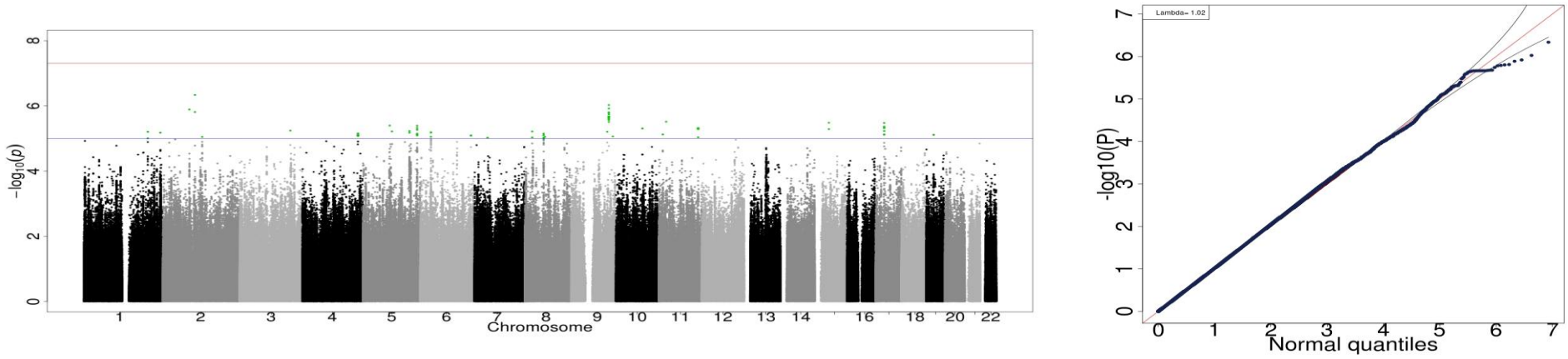


Figure 30: Manhattan Plot and Q-Q plot SUMMIT DN Phenotype Type 1 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	OR_95U	OR	OR_95L	Within Gene
rs11123857	2	1.02E+08	4.61E-07	0.29	G	1.37	1.25	1.14	*NPAS2*
rs11793270	9	1.19E+08	9.43E-07	0.07	T	0.77	0.65	0.55	*LINC00474*
rs62154650	2	84001447	1.29E-06	0.10	A	0.75	0.62	0.52	*NA*
rs79336030	11	22785294	3.06E-06	0.03	A	0.68	0.50	0.37	*GAS2*
rs71478350	15	45263421	3.29E-06	0.03	C	2.33	1.78	1.37	*C15orf43*

Table 45: Meta-Analysis of CKD phenotype in SUMMIT consortium in patients with Type 2 Diabetes

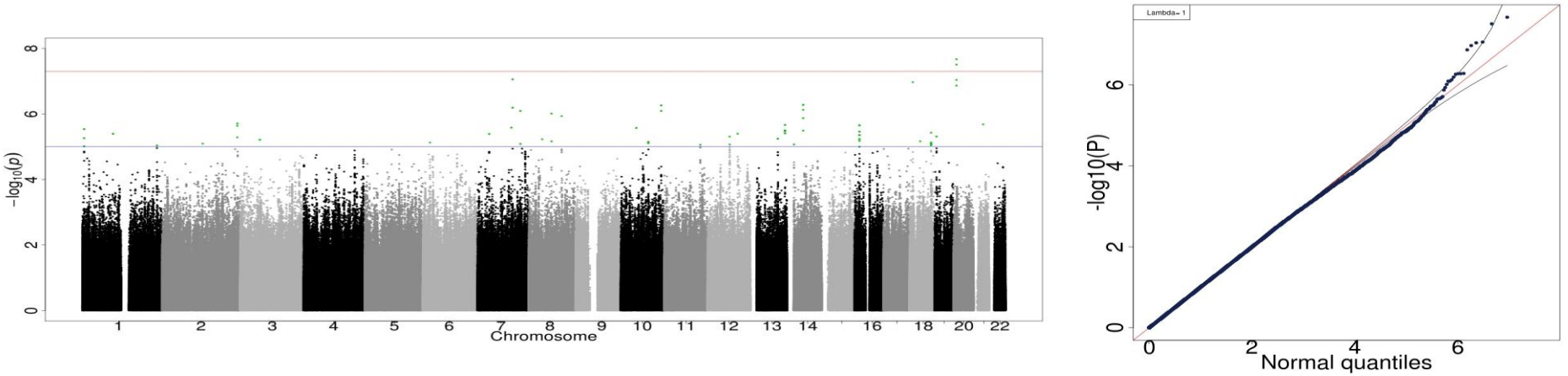
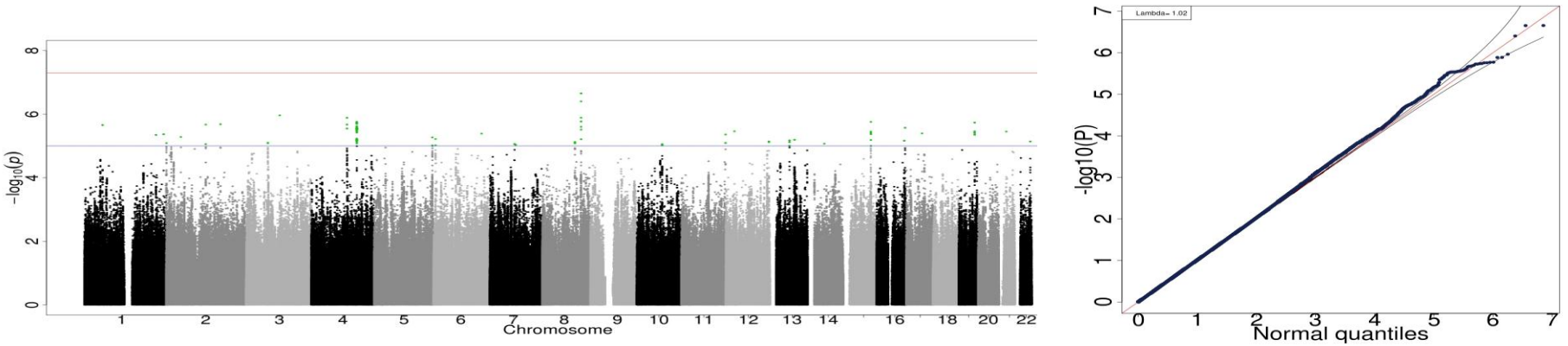


Figure 31: Manhattan Plot and Q-Q plot SUMMIT CKD Phenotype Type 2 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	OR_95U	OR	OR_95L	Within Gene
rs2206136	20	9351150	2.13E-08	0.42	A	1.32	1.20	1.10	*PLCB4*
rs2293280	7	1.09E+08	8.73E-08	0.07	G	0.86	0.72	0.61	*AC073071.1*
rs206439	18	10444655	1.07E-07	0.54	T	1.33	1.21	1.11	*RP11-243E13.1*
rs61981778	14	49648584	5.20E-07	0.05	C	1.96	1.57	1.25	*NA*
rs12251637	10	1.26E+08	5.44E-07	0.04	T	2.07	1.65	1.32	*LHPP*

Table 46: Meta-Analysis of CKD-DN phenotype in SUMMIT consortium in patients with Type 2 Diabetes



QQ-Plot

Figure 32: Manhattan Plot and Q-Q plot SUMMIT CKD-DN Phenotype Type 2 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	other_allele	OR_95U	OR	OR_95L	eaf	Within Gene
rs6997279	8	1.18E+08	2.22E-07	0.25	T	G	1.79	1.51	1.28	0.25	*SLC30A8*
rs188700382	3	1.01E+08	1.09E-06	0.02	A	G	12.07	5.56	2.57	0.02	*CEP97*
rs183409769	4	1.08E+08	1.30E-06	0.12	A	G	2.59	1.95	1.46	0.12	*NA*
rs138532692	15	84637219	1.75E-06	0.18	G	A	0.80	0.66	0.54	0.18	*ADAMTSL3*

Table 47: Meta-Analysis of ESRD phenotype in SUMMIT consortium in patients with Type 2 Diabetes

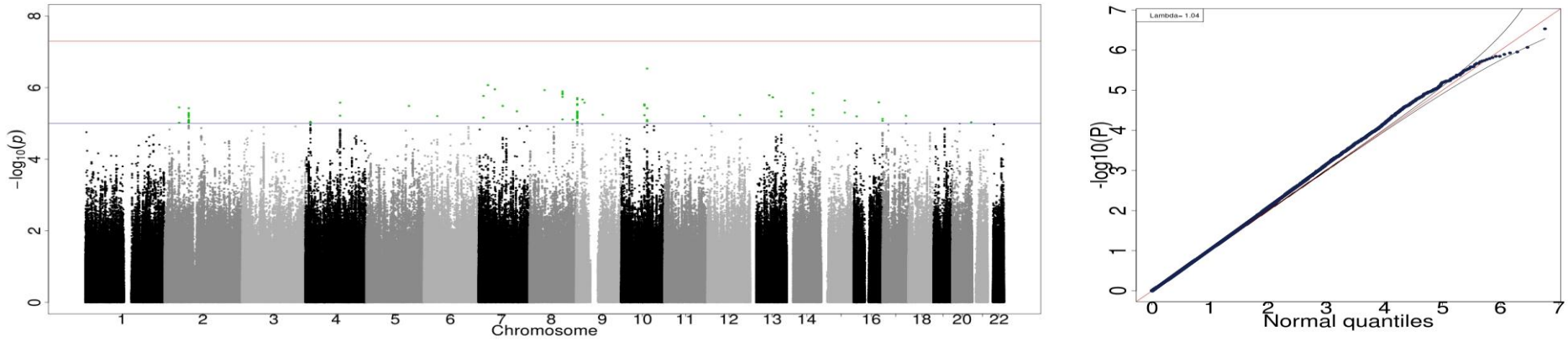


Figure 33: Manhattan Plot and Q-Q plot SUMMIT ESRD Phenotype Type 2 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	other_allele	OR_95U	OR	OR_95L	Within Gene
rs2475363	10	81278814	2.94E-07	0.04	C	T	10.65	4.74	2.11	*NA*
rs13224128	7	29283395	8.54E-07	0.04	T	C	6.27	3.68	2.17	*CHN2*
rs149034284	7	51066779	1.11E-06	0.03	A	G	8.49	4.23	2.11	*RP4-724E13.2*
rs117392950	8	47419544	1.17E-06	0.08	A	T	5.50	3.27	1.94	*NA*

Table 48: Meta-Analysis of Macro-ESRD phenotype in SUMMIT consortium in patients with Type 2 Diabetes

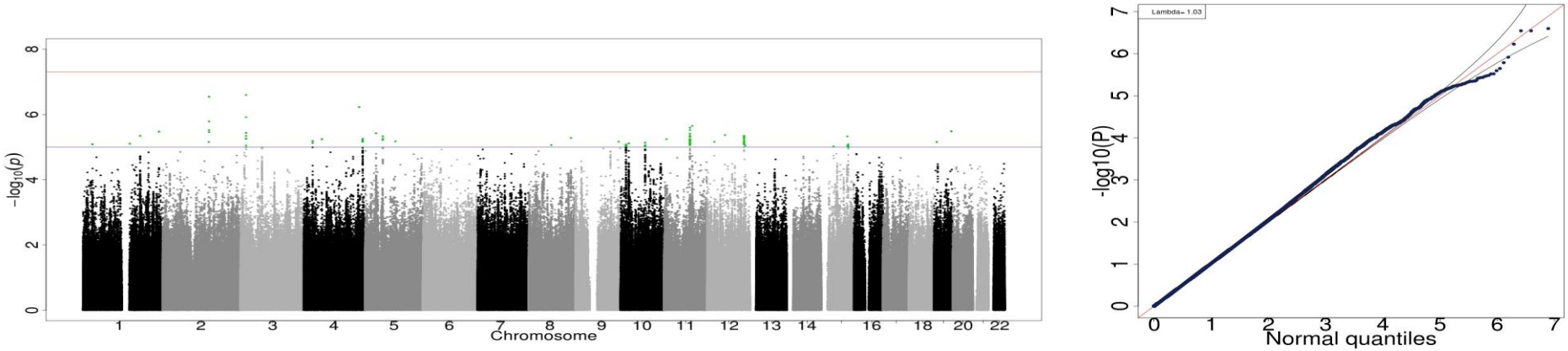


Figure 34: Manhattan Plot and Q-Q plot SUMMIT Macro-ESRD Phenotype Type 2 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	other_allele	OR_95U	OR	OR_95L	Within Gene
rs116354014	3	17170864	2.54E-07	0.02	C	A	8.06	4.52	2.53	*NA*
rs112765093	2	1.45E+08	2.87E-07	0.03	C	G	4.70	2.94	1.84	*GTDC1*
rs147523203	4	1.73E+08	5.95E-07	0.01	A	G	44.53	17.32	6.74	*GALNTL6*
rs7942230	11	88469683	2.25E-06	0.43	T	G	0.86	0.77	0.69	*GRM5*
chr11:81277664:D	11	81277664	2.55E-06	0.47	A	AC	1.48	1.33	1.19	*NA*

Table 49: Meta-Analysis of Micro phenotype in SUMMIT consortium in patients with Type 2 Diabetes

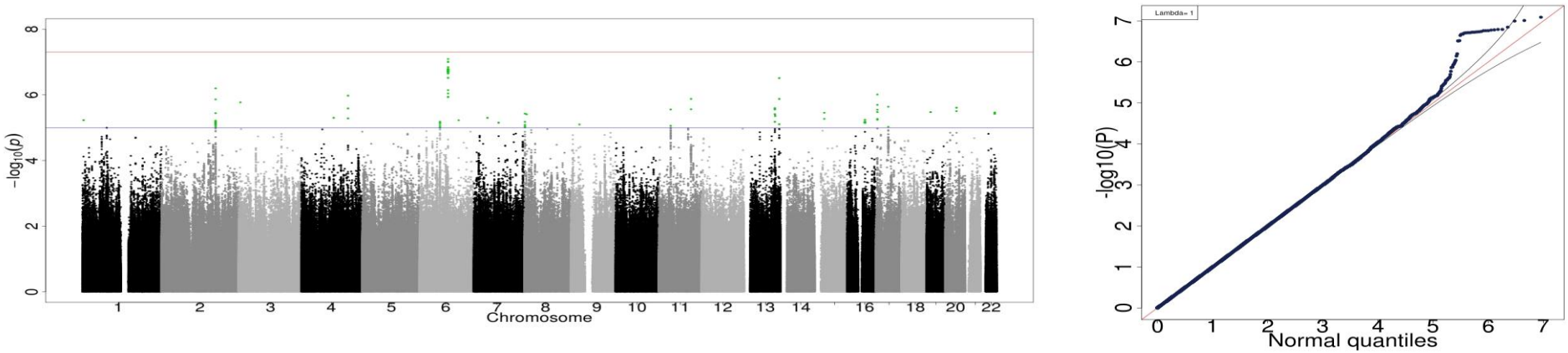


Figure 35: Manhattan Plot and Q-Q plot SUMMIT Micro-albuminuria Phenotype Type 2 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	other_allele	OR_95U	OR	OR_95L	Within Gene
rs2150814	6	89942934	8.10E-08	0.36	T	G	0.88	0.79	0.72	*GABRR1-GABRR2*
rs1143914	13	1.11E+08	3.07E-07	0.49	G	A	0.88	0.80	0.72	*COL4A1*
chr2:170759049:D	2	1.71E+08	6.33E-07	0.53	A	AG	1.40	1.26	1.14	*UBR3*
rs11651114	17	4696736	9.77E-07	0.44	A	G	1.43	1.29	1.16	*NA*
chr4:146704225:D	4	1.47E+08	1.06E-06	0.28	C	CAT	1.47	1.31	1.17	*ZNF827*

Table 50: Meta-Analysis of DN phenotype in SUMMIT consortium in patients with Type 2 Diabetes

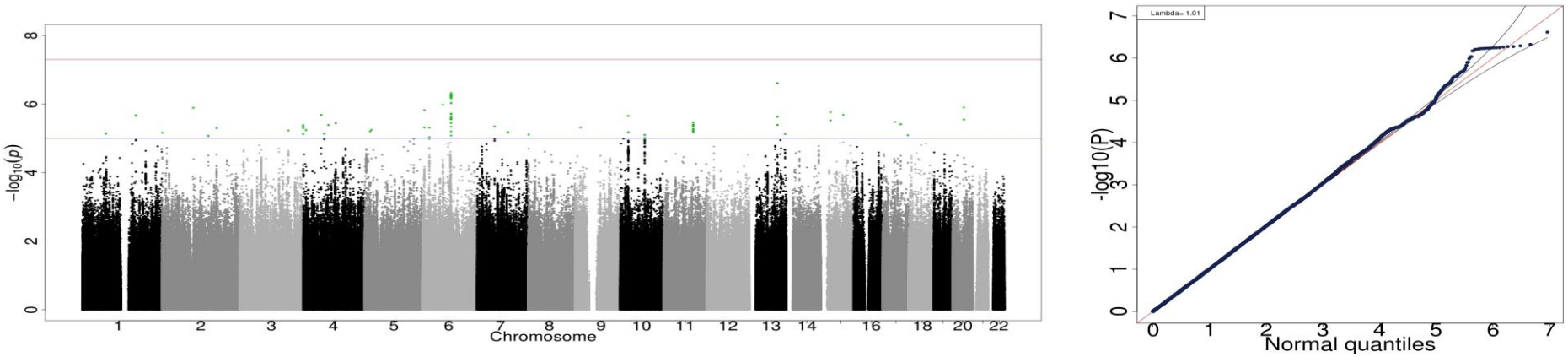


Figure 36: Manhattan Plot and Q-Q plot SUMMIT-DN Phenotype Type 2 Diabetes

RS_NUMBER	CHR	BP	P	eaf	reference_allele	other_allele	OR_95U	OR	OR_95L	Within Gene
rs183249293	13	86971561	2.44E-07	0.02	G	A	0.61	0.42	0.29	*NA*
rs9942471	6	89948232	4.76E-07	0.36	C	A	0.91	0.83	0.76	*NA*
rs2347470	6	63997318	1.04E-06	0.51	A	G	0.90	0.83	0.76	*LGSN*
rs139484064	20	35620979	1.25E-06	0.03	C	T	0.70	0.54	0.41	*NA*
rs182638570	2	97805602	1.28E-06	0.02	T	C	0.44	0.27	0.17	*ANKRD36*

Table 51: Replication of reported loci associated with diabetic kidney disease in SUMMIT meta-analysis

				DN		CKD		MACRO+ESRD		ESRD		MICRO		Directionvs. original finding
Gene	SNP	EA	NEA	OR	P	OR	P	OR	P	OR	P	OR	P	
-	rs1411766	G	A	1.00	0.999	0.99	0.886	1.00	0.911	0.97	0.612	1.00	0.951	
ACACB	rs2268388	G	A	0.98	0.613	0.94	0.258	1.01	0.816	0.98	0.740	1.09	0.156	
ACE	rs1800764	T	C	1.01	0.794	1.07	0.080	1.00	0.945	1.03	0.627	1.08	0.102	
ADIPOQ	rs17300539	G	A	0.88	0.077	0.88	0.122	0.84	0.062	0.73	0.030	0.89	0.238	same
AFF3	rs7583877	T	C	1.01	0.771	0.89	0.002	0.92	0.041	0.78	0.00002	1.06	0.198	same
AGT	rs699	G	A	1.04	0.192	0.98	0.587	1.02	0.549	1.04	0.372	1.10	0.020	same
AGTR1	rs5186	C	A	1.05	0.295	1.01	0.905	1.08	0.161	1.03	0.817	1.01	0.879	
AKRB1	rs3896278	T	C	1.03	0.432	1.00	0.877	1.04	0.372	1.02	0.692	1.08	0.075	
APOC1	rs4420638	G	A	1.05	0.333	1.08	0.143	1.00	0.959	1.01	0.878	1.04	0.567	
CARS	rs739401	T	C	1.04	0.204	0.98	0.579	1.04	0.313	1.02	0.659	1.03	0.474	
CARS	rs451041	G	A	1.04	0.197	0.98	0.582	1.06	0.308	1.02	0.702	1.03	0.502	
CPVL/CHN2	rs39075	G	A	1.00	0.988	0.93	0.053	1.00	0.950	0.98	0.692	1.00	0.913	
CPVL/CHN2	rs39059	G	A	1.01	0.874	1.05	0.175	0.99	0.746	1.02	0.683	1.03	0.478	
ELMO1	rs1558688	T	C	1.04	0.213	1.06	0.140	1.04	0.290	1.18	0.004	1.01	0.851	Opposite to JPT. same with European (12)
EPO	rs1617640	C	A	1.08	0.012	1.02	0.700	1.03	0.509	0.99	0.803	1.08	0.085	same
ERBB4	rs7588550	G	A	0.86	0.042	0.91	0.281	0.80	0.015	0.78	0.062	0.87	0.191	same
FRMD3	rs942280	T	C	1.00	0.883	1.04	0.282	1.05	0.266	1.01	0.847	0.99	0.750	
FRMD3	rs10868025	G	A	1.00	0.932	0.98	0.570	0.98	0.692	0.93	0.211	1.08	0.068	
FRMD3	rs1888747	G	C	0.99	0.720	0.98	0.687	1.00	1.000	1.08	0.173	0.92	0.057	
GLUT1	rs841853	C	A	1.03	0.435	0.99	0.751	1.02	0.615	0.99	0.892	1.01	0.846	
GREM1	rs1129456	T	A	1.05	0.348	1.01	0.903	1.01	0.072	1.06	0.519	0.99	0.886	
HSPG2	rs3767139	T	C	1.04	0.303	0.99	0.743	1.02	0.622	1.02	0.745	1.03	0.587	
LIMK2	rs2106294	T	C	1.03	0.464	1.01	0.832	1.08	0.074	1.03	0.656	0.99	0.751	
LOC100132891	rs9298190	T	C	0.94	0.041	0.97	0.384	0.93	0.062	0.95	0.332	0.93	0.111	same
MSRB3-HMGA2	rs2358944	G	A	1.02	0.656	1.00	0.929	0.98	0.716	0.98	0.731	1.02	0.692	

NOS3	rs1800779	G	A	0.96	0.230	0.92	0.039	0.97	0.104	0.96	0.466	0.98	0.683	same
PPARG	rs1801282	G	C	1.06	0.201	1.11	0.041	1.05	0.342	1.05	0.497	0.98	0.696	same
PRKGA2	rs7805747	G	A	1.01	0.764	0.94	0.362	1.05	0.334	0.99	0.872	0.99	0.872	
PVT1	rs11993333	T	C	0.98	0.506	0.98	0.542	0.98	0.560	1.01	0.888	0.97	0.408	
PVT1	rs2648875	G	A	1.03	0.353	1.02	0.567	1.03	0.523	1.03	0.623	0.99	0.792	
RGMA – MCTP2	rs12437854	T	G	0.85	0.066	0.85	0.111	1.04	0.732	0.77	0.092	1.03	0.801	
RPS12	rs7769051	C	A	0.99	0.892	0.94	0.279	0.98	0.736	0.99	0.226	1.00	0.965	
SASH1	rs6930576	G	A	0.97	0.314	0.95	0.165	0.93	0.089	0.96	0.472	0.98	0.582	
SOX11	rs16864170	T	C	0.98	0.844	0.97	0.791	0.93	0.523	1.03	0.865	0.95	0.703	
UMOD	rs12917707	T	G	0.98	0.558	0.94	0.213	1.00	0.937	1.01	0.873	0.99	0.780	
UNC13B	rs13293564	T	G	1.02	0.437	1.04	0.333	1.07	0.097	1.02	0.694	1.00	0.975	
VEGFA	rs833061	T	C	1.01	0.690	0.98	0.597	1.02	0.595	1.03	0.650	0.96	0.404	
ZMIZ1	rs1749824	C	A	1.02	0.569	0.99	0.843	1.02	0.608	1.03	0.597	0.99	0.887	

Discussion

I. Estimating chip-based heritability of various DN phenotypes

A key challenge in performing GWAS for DKD that based on albuminuria and eGFR levels, there are many ways in which the phenotype can be defined (micro-albuminuria, macro-albuminuria, ESRD, CKD, and CKD with and without Albuminuria). At the onset, some prior information about the heritability sub-phenotypes might help in the prioritization of the DKD phenotypes. Accordingly, we calculated chip-based heritability of various DKD sub-phenotypes and aimed to understand if there are differences in the chip heritability of various sub-phenotypes of DKD.

Genome complex trait analysis (GCTA) [154] [121] has been used to estimate chip-based heritability for several common complex disorders, including diabetes, hypertension and hyperlipidaemia. The primary motivation for developing GCTA was to identify the "hidden heritability" for complex phenotypes which the GWAS studies failed to explain. GWAS studies usually report estimates from single SNP association analysis and may not have adequate power and effect sizes to detect SNPs with modest effects, given the rigorous threshold of p-values and replications. Hence, the method measures the additive genetic variance explained by all the SNPs combined in the GWAS study on the background of genetic relation matrix estimated from GWAS SNPs[121]. This could be useful approach as it can provide *a priori* information about heritability of a trait before running a GWAS which in turn can help in prioritizing phenotypes. Another function of GCTA is the novel bivariate analysis. This approach has advantages over the approach described above as it can be used to identify shared genetic determinants between two correlated states or

phenotypes estimated by genetic correlation r_g. An $rg_{=}1$ will imply that there is a complete overlap in genetic variants shared between two states or phenotypes while an $r_{g<}1$ will mean that at-least some genetic variants are underlying the two traits are different. This is important to understand if there are shared genetic determinants of co-related phenotypes such as blood pressure and eGFR or Hba1C and eGFR.

One of the major drawbacks of GCTA is the fact that it ignores heritability due to factors such as gene-gene interactions, copy number variations etc. Thus, the estimates represent the lower bounds of the actual heritability- estimated from the genome-wide representative SNPs giving rise to the term "chip based heritability". This is evident from the results of GCTA, for example, human height is about 80% heritable; however, the estimates from GCTA for several studies are around 40%[121].

In the present study, we categorized the cases into different subgroups based mainly on albuminuria excretion and presence and absence of CKD. The phenotypes interrogated for their chip-based heritability were eGFR as a continuous trait, eGFR as a categorical variable (eGFR<60) a combination of albuminuria and eGFR (CKD-DN) (eGFR<45 and macro-albuminuria vs eGFR>60 and no albuminuria), macro-albuminuria, micro-albuminuria and any albuminuria. Our study shows some evidence that CKD, CKD-DN, macro-albuminuria, and ESRD are more heritable as compared to micro-albuminuria and any-albuminuria phenotype. This is consistent with the observation that micro-albuminuria is a more variable trait and is more likely to be influenced by environmental factors such as fever and transient hyperglycaemia[155]. This can decrease the power to detect variations in micro-albuminuria due to genetic determinants. However, the confidence intervals associated with the chip-based heritability estimates were very high precluding us from making any meaningful conclusions.

We conducted a bivariate analysis of blood pressure and eGFR. Our study showed limited overlap of genetic determinants underlying eGFR and blood pressure suggesting that there is a possibility of identifying variants affecting eGFR, which do not act through their effect on blood pressure. Again, the confidence intervals associated with the estimates genetic-correlation between the two phenotypes were very high precluding us from making any meaningful conclusions.

We calculated the chip-based heritability with both imputed and directly genotyped data. Interestingly, the difference between the heritability estimates calculated from imputed and directly genotyped data was not significant. The use of imputed data did not show a significant increase in heritability, even with a substantial increase in the number of interrogated variants, suggesting that the directly genotyped data alone sufficiently captured the chip-based heritability present in SNP level data.

A key limitation of this study is the relatively small sample size studied. It would be interesting to combine individual level GWAS data across all the SUMMIT cohorts and see its effect on heritability estimates for various albuminuria phenotypes. However, combining genotype data across different cohorts and platforms is a challenging task because of a) governance challenges in acquiring individual level data and b) differences in genotyping due to use of different genotyping platforms.

In summary we showed some suggestive evidence that macroalbuminuria, ESRD and chronic kidney disease (eGFR<60) are be more heritable than other DKD phenotypes; however, because of small sample sizes and limited power we could not prioritise the GWAS phenotypes based on their chip-based heritability.

II. GWAS for upstream risk factors of DKD (BP and HbA1c) OR doing Genetic risk scores with known loci and testing their association with DN phenotypes

In this study, we provide first replication of Blood Pressure and HbA1C associated SNPs in patients with Type 2 diabetes (n=~7000) and study the cumulative effect of known BP and HbA1C associated SNPs on diabetic kidney disease.

There is an increasing interest in investigating the role of abdominal obesity, dyslipidaemia, hypertension, fasting hyperglycaemia and HbA1c in the development of kidney disease in patients with diabetes[156]. Several studies have shown association of higher HbA1c and high blood pressure with onset of kidney disease in patients with Type 2 Diabetes. Multiple large-scale clinical trials, including the Diabetes Complications and Control Trial (DCCT) in Type 1 diabetes[157] and the United Kingdom Prospective Diabetes Study (UKPDS)[158] in Type 2 diabetes, have shown that the risk for DKD begins to increase at an HbA1c level of 6.5%. Hypertension on the other hand has long been recognized as both, a consequence of renal impairment and an important factor in the progression of DKD[156]. Shulman et al showed that the incidence of a decline in renal function over 5 years is greater among older patients with hypertension[159] and Retnakaran et al showed that hypertension is an independent risk factor for development of albuminuria or renal impairment among patients with type II diabetes[160]. Since hypertension is both a consequence and a risk factor of DKD, a conclusive role of hypertension as an upstream risk factor for DKD a can be established by Mendelian randomization study looking at the association for genetic determinants for high blood pressure and DKD. Similarly, an association between the genetic risk score for high HbA1C and

DKD will clearly establish HbA1C as a risk factor for DKD. Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in non-experimental studies[161]. Gray and Wheatley (1991)[161] first described this method for obtaining unbiased estimates of the effects of a putative causal variable without conducting a traditional randomised trial.

There have been several genome-wide association studies for hypertension. The HYPERGENES Project was a two-staged case controlled study by Salvi et al showed that significant association exists between HTN and rs3918226 located on eNOS gene in its promoter region[162]. Newton et al conducted a GWAS and identified eight loci associated with HTN[163]. The loci identified were *CYP17A1, CYP1A2, FGF5, SH2B3, MTHFR, c10orf107, ZNF652* and *PLCD3*. Padmanabhan et al conducted a GWAS nearly 1600 HTN patients and 1700 controls and showed that a locus on chromosome 16, rs13333226, in the uromodulin gene region was identified to have significant association with HTN[164]. In a joint meta-analysis of 30,000 individuals in CHARGE Consortium (Cohorts for Heart and Aging Research in Genome Epidemiology) with Global BP gen consortium (n=35,000 patients) and 10 novel SNPs associated with BP were identified[165]. GWAS significance was attained in four (*ATP2B1, CYP17A1, PLEKHA7, SH2B3*) SNPs for SBP, 6 SNPs (*ATP2B1, CACNB2, CSK/ULK3, SH2B3, TBX3/TBX5, ULK4*), for DBP.

Percent HbA1c (glycated haemoglobin) is an informative trait for diabetes diagnosis and management and is used in quantifying chronic glycaemic exposure of erythrocytes for the preceding 2 to 3 month. There is strong correlation between HbA1c levels and occurrence of diabetes complications. There are several GWAS studies looking at genetic determinants of HbA1c levels. In 2010, MAGIC (Meta-

analyses of Glucose and Insulin related traits Consortium) identified ten genetic loci associated with HbA1c[166]. The ten loci associated included three loci in or near genes likely involved in glycaemic control pathways: *G6PC2, GCK*, and *MTNR1B* and seven loci in or near genes likely to be involved in erythrocyte biology, including *SPTA1, HFE, ANK1, HK1, APT11A, FN3K,* and *TMPRSS6*. A large study by Paré et al performed in 14,618 non-diabetic individuals identified *HK1* (encodes for enzyme hexokinase) as associated with HbA1c levels[167]. Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group (DCCT-EDIC) have reported *SORCS1* as major locus for HbA1c levels in patients with Type 1 Diabetes[168].

The studies identifying SNPs for both hypertension and HbA1c have been performed in general population with a small proportion of patients with Type 2 diabetes. In fact, the biggest study identifying genetic determinants of HbA1c (MAGIC) was performed in non-diabetic participants. Type 2 Diabetes and concomitant presence of DKD can either attenuate or exaggerate the effects of these SNPs on hypertension and HbA1c hence it is important to test the effect of known HbA1c loci in patients with diabetes. Hence, we decided to perform a replication analysis of the top GWAS SNPs associated with blood pressure and HbA1c in patients with Type 2 diabetes. We identified SNPs convincingly associated by blood pressures and HbA1c from the Catalogue of Published Genome-Wide Association Studies[169] and selected 60 SNPs associated with blood pressures and 14 SNPs associated with HbA1c. Of the 60 SNPs for hypertension, 20 were significantly associated with blood pressure after correction of multiple testing; while from 14 SNPs associated with HbA1c eight SNPs from six genes were associated with Hba1c in the Type 2 diabetes cohort.

This suggests that the genetic determinants of blood pressure and HbA1c in Type 2 diabetes are shared with those in general populations.

Next, we looked at the effect of the variants associated with HbA1c and blood pressure on albuminuria in patients with Type 2 diabetes. Since these variants explain a small proportion of (usually less than 1%) heritability of hypertension and HbA1c[166], we decided to study the cumulative effect of these variants on DKD and calculated beta-weighted genetic risk scores for each study participant. The age and sex adjusted genetic risk scores for hypertension and HbA1c were significantly associated with macro-albuminuria and ESRD phenotypes in the Go-DARTS datasets. This Mendelian randomization approach conclusively establishes the causal effect of hypertension and HbA1c on DKD. We show that genetic risk score for blood pressure is associated with DKD leading us to presume that BP is upstream risk factor of DKD. However, it is possible that some of these loci have an effect on the physiology, biochemistry of kidneys which in turn regulate the blood pressure. Thus, the effect of these SNPs can be on kidneys rather than the blood pressure itself. However, as most of the SNPs have been associated with blood pressure in non-diabetic non- kidney disease population , it can be assumed that majority these SNPS act of the kidneys by their action on blood pressure and not *vice versa*.

In summary, we replicated the association of several known loci for Hypertension and HbA1c in patients with Type 2 diabetes and show that these variants have a cumulative effect on albuminuria in patients with Type 2 diabetes.

III. VI. Replication of CKDgen Consortium eGFR loci Hits

We replicate the association of *UMOD GCKR*, and *SHROOM3* with eGFR in patients with Type 2 diabetes. Our study confirms the findings of previous studies showing the association of UMOD with eGFR and diabetic nephropathy[170-173] and association of GCKR, and SHROOM3 with eGFR[174-176]. A study by Gudbjartsson et al showed and interaction of UMOD with age[173]; while another study[170] could not replicate this interaction. In the present study we do not see an interaction of UMOD with age in a patients with Type 2 Diabetes (P=0.84).

None of the other variants were associated with eGFR after correction for multiple testing; however, the direction of effect for most of the SNPs was consistent with the previous studies. Our study had limited power to estimate the effect of these variants on eGFR. Taken together all these variants explain 1.4-14% heritability of eGFR[177] (with each SNP contributing typically less than 0.5% heritability of eGFR). Our study had 97% power to detect an association with a SNP explaining 0.5% variability in eGFR due to individual SNP and anything below 0.5% can remain undetected. It is also possible that some of these SNPs are not the causal SNPs and because of varying linkage disequilibrium, (LD) structure in our population could not be detected. It is also possible that effects of some of these SNPs were attenuated by diabetes or diabetic kidney disease and hence were not associated with eGFR in this study.

We examined the association of these loci with decline in renal function using a Cox-proportional hazard model and estimated effect of these loci on time to stage 3B CKD (eGFR<45mls/min/1.73m2). Given the high mortality associated with diabetic nephropathy, cross-sectional studies are prone to survival bias, as patients with

severe forms of nephropathy are less likely to be included in the study. Hence, it is important to investigate the eGFR loci in a time dependent manner. Of the 15 SNPs, only rs12917707 in *UMOD* was associated with time to stage 3B-CKD with the minor allele 'T' showing a protective effect. *UMOD* has a stronger effect on baseline eGFR as compared to other 14 markers suggesting that SNPs with a strong effect on baseline eGFR influences decline in renal function over a period.

We did a stratified analysis to examine the effect of albuminuria on the known genetic associations with eGFR. In T2D, nephropathy albuminuria may be more closely associated with decline in renal function and the impact of genetic determinants of eGFR may differ depending on the presence or absence of nephropathy. Hence, we examined the effects on eGFR stratified by albuminuria. There is clear difference in the effect sizes in those with sustained normalbuminuria and those with albuminuria. For example, the *UMOD* has twice the effect in patients with sustained normalbuminuria as compared to those with albuminuria (P-interaction=0.002) while *SHROOM3* (P-interaction=0.003) and *GCKR* (P-interaction=0.08) had larger effect sizes in those with albuminuria. It is known that kidney diseases characterized by albuminuria, such as diabetic nephropathy can have ultrafiltration and high eGFR in the early stage of disease while those characterised by reduced renal function like hypertensive kidney disease may manifest with normo-albuminuria because of the reduced renal efficiency[71,72]. Hence, studying the genetic determinants of eGFR without adjusting for albuminuria status or studying genetic determinants of albuminuria without accounting for eGFR can reduce the power of these studies to identify true genetic effects. Cumulatively, eGFR associated loci explain only a small fraction of the total heritable contribution eGFR

and stratifying by albuminuria status in our existing GWAS datasets[177-179] can help us to uncover the missing heritability. The heterogeneity in the effect sizes suggests different mechanisms by which these genes impact upon the kidney function; it is possible that *UMOD* is a basal kidney function gene and the diabetic albuminuria attenuates its effect on renal function. It is possible that the heterogeneity seen at the *UMOD, SHROOM3 and GCKR* arises due to difference in the baseline characteristics *between* the two cohorts (i.e with (n=2097) and without albuminuria (n=613)). However, the regression analysis adjusted for baseline characteristics such as age, sex, BMI, BP and duration of diabetes thereby minimizing the effects of baseline differences

In summary, our results show that some of the genetic determinants of eGFR in general population are common to patient with Type 2 diabetes. However, in patients with Type 2 diabetes it is essential to adjust for albuminuria status while investigating the genetic determinants of renal function.

IV. GWAS for biomarker for DN (sRAGE)

We report the first genome-wide association study to identify genetic determinants of circulating sRAGE levels and confirm the association of rs2070600 (G28S), a non-synonymous SNP located in exon 3 of RAGE, with circulating sRAGE and esRAGE levels. Moreover, we identified novel polymorphisms rs10940285 in *ITGA1-ITGA2* gene region and rs9272346 in *AGER-HLA-DQA1* associated with sRAGE levels.

Association of known variant in AGER gene with sRAGE levels

The 'T' allele in rs2070600 (G28S) encodes for serine and is shown to be associated with lower sRAGE levels in European[180,181] and Asian populations[181] and our study confirms this findings. Mechanisms whereby SNP rs2070600 (G28S) affects the sRAGE plasma concentrations remain unknown. It has been proposed that the alteration of the N-glycosylation state of the protein, caused by the rs2070600 (G28S) polymorphism, induces structural changes in the protein that makes RAGE more vulnerable for the action of proteinases, such as a disintegrin and metalloproteinases 10[182] and matrix metalloproteinase-9[183]. The minor allele 'T' of rs2070600 (associated with lower sRAGE levels) has been associated with increased risk of Alzheimer's disease[184], Type 1 diabetes[185] and asthma[186]. All three diseases have been associated with lower sRAGE levels[184]. Two other SNPs of *AGER*, previously associated with HLA-DR/DQ haplotype independent diabetes risk (rs9469089, and rs17493811)[187], were not associated with sRAGE levels in this study.

Association of HLA-DQA1 gene (HLA Class II region) with sRAGE levels

We identified a novel marker rs9272346 in the HLA-DQA1 gene associated with the sRAGE levels. A genome-wide conditional analysis using summary statistics from the CARDS- Go-DARTS meta-analysis (with the algorithm implemented in GCTA) suggests that the association of rs9272346 with the sRAGE levels is independent of the known marker rs2070600 (G28S) in the *AGER* gene. The 'G' of rs9272346 allele was associated with higher sRAGE levels. Carriers of "G" allele of rs9272347 have lower risk for T1D and Asthma [188,189] and interestingly, sRAGE is deficient in both T1D and in neutrophilic asthma and chronic obstructive pulmonary disease

(COPD)[190,191]. Although this variant is independent of the known rs2070600 *(G28S)* in the *AGER* gene, it is possible that this variant could be in long range LD with another SNP in *AGER* gene.

Association of ITGA1 gene with sRAGE levels

We identified a novel marker rs10940285 in *3 prime UTR* of *ITGA1* gene associated with sRAGE levels. This gene encodes the alpha 1 subunit of integrin receptors. This protein heterodimerizes with the beta 1 subunit to form a cell-surface receptor for collagen and laminin. The heterodimeric receptor is involved in cell-cell adhesion and may play a role in inflammation and fibrosis. The alpha 1 subunit contains an inserted (I) von Willebrand factor type I domain which is thought to be involved in collagen binding. *ITGA1* has demonstrated genetic pleiotrophy as variants in the *ITGA1* have been associated with bone mineral density, Type 2 diabetes, fasting insulin, β-cell function by homeostasis model assessment, and 2-h post–oral glucose tolerance test glucose and insulin levels, liver fibrosis, insulin secretion, and bone healing[192-195]. While it is unclear how *ITGA1* affects sRAGE levels, we hypothesize that through its action on fasting glucose and insulin, *ITGA1* might affect the formation of advanced glycation end-products which in turn can regulate the sRAGE levels. It is interesting to note that the effect of this polymorphism on sRAGE levels is more significant in patients with Type 2 diabetes. We did not find association between rs10940285 in *ITGA1* and fasting glucose and insulin in the CARDS study, however, this analysis might be limited by the smaller sample size as compared to original reports. Interestingly, the RAGE ligands include matrix proteins such as Collagen I and IV[196]. *ITGA1* forms cell-surface receptors for

collagen n and through ligand, sharing by *ITGA1* might influence the levels of circulating sRAGE.

Association of promoter variants in AGER gene with sRAGE levels

Two promoter variants in the *AGER* gene rs1800624 (374 T>A) and rs1800625 (429 T>C) have been frequently investigated for their association with diabetes, cardiovascular disease and other metabolic traits. For example, rs1800624 (374T>A) is shown to be associated cerebrovascular disease in both Type 1 and Type 2 diabetes as well as in those without diabetes[197-200], while rs1800625 has been associated with pre-diabetic state, with higher HbA1c and insulin resistance[201,202]. Neither rs1800624 (P-value=0.59) nor rs1800625 (P-value=0.22) were associated with sRAGE level the CARDS-Go-DARTS meta-analysis.

Chip-based heritability for sRAGE levels

We provide first evidence for the heritability of sRAGE levels using GCTA. One drawback of GCTA is the fact that it ignores heritability due to factors such as, copy number variations, runs of homozygosity and other sources of heritability[148]. Thus, the estimates represent the lower bounds of the real heritability, giving rise to the term "chip-based heritability". Nonetheless, we show that the point-estimate for chip-based heritability of sRAGE is 37% and the three genome-wide significant SNPs in the study explain 10% of this heritability in European populations.

GWAS for sRAGE/esRAGE ratio

We performed a GWAS for splice variant endogenous secretory RAGE (esRAGE) levels and esRAGE/sRAGE ratio using the GWAS data in the CARDS dataset. We did not find any variant differentially associated with esRAGE as compared to sRAGE levels and no marker reached the accepted threshold of significance in the GWAS for esRAGE/sRAGE ratio (data not shown). It has been shown that esRAGE constitutes only a small proportion of the total sRAGE pool and is highly correlated with the sRAGE levels[203]. Since the esRAGE and sRAGE are highly correlated, the identification of genetic variants specifically associated with esRAGE and not with sRAGE will need study with very large sample size. There is very limited epidemiological data about esRAGE/sRAGE ratio, with some reports indicating that the ratio can change in diseases states like pre-eclampsia[204]. Further epidemiological characterization of the esRAGE/sRAGE ratio will be useful to identify the genetic determinants of the esRAGE/sRAGE ratio.

Study Limitations

Our study had several limitations. With a discovery, sample-size of 1,377 it had 60% power to detect a SNP, which explains 2% variation in sRAGE levels, and only 9% power to detect a SNP, which explains 1% variation in sRAGE levels. Hence, variants with small effects on sRAGE level could have been missed in this GWAS analysis. Additionally, since the GWAS chips and imputations capture only the common tagging variants in the genome, the novel variant rs9272346 could be tagging a causal underlying variant and a dense fine-mapping and sequencing of this region might help to identify this variant.

In summary, we identified in novel variants in the in *ITGA1-ITGA2* and *HLA-DQA1* region associated with sRAGE levels, confirmed the association of rs2070600 (G28S) with sRAGE and show that sRAGE levels have a heritable component. Further studies are needed to investigate the mechanisms by which *ITGA1-ITGA2* and *HLA-DQA1* influence *sRAGE* levels. Adequately powered GWAS studies can identify the additional variants associated with sRAGE. The identified variants can be utilized in Mendelian randomization studies to establish the causality between sRAGE and diabetic complications and vascular diseases, and further functional studies of the variants can identify pathways involved in the AGE-RAGE axis and vascular diseases.

V. Meta-analysis of GWAS data with other DKD datasets

We performed a genome-wide association study to investigate genetic susceptibility to Diabetic Kidney Disease in Type 1 and Type 2 diabetes. Previous genome-wide association studies[97,112,114,115,118,119] for diabetes-associated kidney disease in Japanese, Pima Indians, and Caucasian Americans have identified several potential DKD loci such as *ELMO1, PVT1, FRMD3*, and *CARS*. As discussed earlier, these GWAS's had several limitations such as a small sample size, varying definitions of DKD and inadequate duration of diabetes for the controls. The proven heritability of DKD using both conventional family based methods and chip-based methods, suggests that GWAS studies should be performed to identify novel genetic determinants of DKD. Hence, we performed GWAS for diabetic kidney disease using for both Type 1 and Type 2 diabetic kidney disease using ~3,000 individuals from the Go-DARTS cohort for Type 2 diabetes and 1,000 individuals for the

EURODIAB cohort for Type 1 Diabetes and meta-analysed these dataset in an international consortium (SUMMIT).

In the results presented below, I performed the GWAS analysis of two Go-darts datasets and EURODIAB dataset for all the DKD phenotypes. **I then assisted a team of analysts in the SUMMIT consortium to perform the GWAS meta-analysis for all the phenotypes.**

Novel Findings for DKD in Type 1 and Type 2 Diabetes

We report two genome-wide significant SNPs for the CKD and Micro-albuminuria phenotype in Type 2 diabetes. For CKD we identified a locus in *PLCB4*, which encodes phospholipase C beta 4. This enzyme catalyses the formation of inositol 1,4,5-trisphosphate and diacylglycerol from phosphatidylinositol 4,5-bisphosphate. This reaction plays an important role in the intracellular transduction of many extracellular signals in the retina and *PLCB4* might be involved in maintenance of the circadian rhythm. Neutrophils are the most abundant subtype of white blood cells (WBCs), comprising 50–70% of all WBCs. Yukinori Okada Hum et.al[205] did a GWAS in 5771 individuals with Japanese ancestry. They identified locus in 20p12 associated with neutrophil count (rs2072910 in PLCB4 at 20p12, $P = 3.1 \times 10{-}10$). This locus is not in LD with the SNP associated with DKD in SUMMIT GWAS analysis. No study has yet shown association of neutrophil count with DKD however, high neutrophil count is known to be associated with lower serum albumin, elevated creatinine, and increased mortality risk in haemodialysis patients[206]. Hence, if this locus were to replicate we can study the association of neutrophil count with DKD in Go-DARTS dataset. (Neutrophil percentage is available in GO-Darts dataset).

For micro-albuminuria, a GWAS significant locus was seen at a SNP between *GABRR1* and *GABRR2*. GABA is the major inhibitory neurotransmitter in the mammalian brain where it acts at GABA receptors, which are ligand-gated chloride channels. *GABRR1* is a member of the rho subunit family. Several transcript variants encoding different isoforms have been found for this gene. *GABRR1* is a known locus for bipolar schizoaffective disorder[207] and alcohol dependence[208]. There is no known association of this locus with diabetes or any other autoimmune disorder. GABRR1/2 is shown to be putatively associated with diabetic cataract in a study performed in Taiwanese population[209]. This study was performed in 109 T2D patients with cataract and 649 patients with no cataract. However, the association of GABRR1/2 with cataract was not genome-wide statistically significant and hence larger studies will be required to confirm this association. The initiating mechanism in diabetic cataract seems is the generation of polyols from glucose by aldose reductase and these can also play role in pathogenesis of DKD . If this locus is successfully replicated for DKD it would be interesting to further study its association with diabetic cataract[210].

We identified three loci were associated with micro-albuminuria at $p < 10^{-7}$, including rs1143914 in the intron of *COL4A1*. This gene encodes collagen type IV, one of the major components of the glomerular basement membrane[211], making it a strong candidate for involvement in DKD. Another interesting locus for Type 2 DN phenotype was chr4:55939605: I ($p=2.1x10^{-6}$) situated near *KDR* gene. *KDR* encodes the *VEGF*-receptor 2 that is expressed on endothelial cells, podocytes and tubular cells, and is known to be essential for normal glomerular function[212]. Both *KDR* and *VEGF* are up-regulated in diabetes[213] and overexpression of *VEGF* in animal models results in proteinuria[214]. For ESRD in Type 2 diabetes a less

significant but biologically relevant signal was detected near *KLF10*. *KLF10* is a transcriptional repressor that acts as an effector of transforming growth factor beta (*TGF-beta)* signalling[215] which is known to be a pivotal player in the development of DKD.

For The T1D- CKD-DN phenotype, we observed a genome-wide significant locus at *AFF3* gene. *AFF3* (AF4/FMR2 family, member 3) has been earlier associated with T1D-ESRD in a meta-analysis including both FinnDiane and SDR[97]. Variants near this gene have been associated with T1D and rheumatoid arthritis[216]. Since, FinnDiane and SDR cohorts have contributed data to this meta-analysis this locus cannot be regarded as a novel locus. For the T1D-ESRD phenotype a strong association was also seen for rs76729345 in *INSR* (Insulin Receptor gene) (P=4.40x10^{-6}). Binding of insulin to the insulin receptor stimulates glucose uptake into cells. *INSR* is expressed in kidneys, and the gene expression is higher in patients with T1D-DKD compared with healthy kidney donors[217]. Variants in this gene have been suggestively associated with diabetic retinopathy in patients with T1D. Nearly all T1D patients with ESRD have at least some level of diabetic retinopathy, and this locus may be a common associated SNP for the two major microvascular complications of T1D. For the macro-albuminuria ESRD phenotype in T1D, we observed a suggestive GWAS significance at *INSIG2*. *INSIG2* encodes an endoplasmic reticulum protein that regulates lipid synthesis by blocking the proteolytic activation of sterol regulatory element binding proteins (SREBPs) and their transportation to the Golgi. Given its role in lipid metabolism, this is an interesting candidate for DKD.

Replication of Previous Results

We studied previous GWAS loci reported in diabetic nephropathy, either T1D or T2D. We also included a number of candidate gene loci that were significantly associated with DN in a literature-based meta-analysis. Further, we included three SNPs significantly associated with CKD in non-diabetic CKD. None of the SNPs were significantly associated with disease after correction for multiple testing. Nominal significance ($P<0.05$) with an effect in the same direction as the original finding was seen for SNPs in *ADIPOQ ELMO1 NOS3 PPARG*. The non-replication of previously reported loci in the GWAS suggests either 1) publication bias in the candidate gene studies whereby only positive findings are reported, or 2) difference in the linkage disequilibrium (LD) pattern between our GWAS cohorts and earlier studies. Although this is not very likely given most of the GWAS and candidate gene studies are reported in European populations, it would mean that the underlying tagging variant could not be captured in our study. 3) Alternatively, it could mean that the underlying variant has a modest effect and our data did not have adequate power to capture it.

Limitations

We will now discuss some of the limitations of the Genome-wide DKD Meta-analysis in the SUMMIT consortium

Heterogeneity of DKD phenotype

Despite of our large sample size, we did not achieve genome-wide statistical significance for our main DN phenotype (any albuminuria vs No Albuminuria with 15-year duration of diabetes) for both Type 1 and Type 2 Diabetes suggesting that

this phenotype may have been too heterogeneous to detect significant associations with a sample of this size. Using the most severe forms of DN to define cases (e.g. ESRD or eGFR<45+macroalbumiuria) reduces some potential misclassification but definitely does not overcome the critical imprecision of case definition not based on histology (many other kidney diseases can cause macro-albuminuria and ESRD).

Cross-sectional study design

The problems associated with cross-sectional study design to identify the genes for diabetes complications and risk factors for diabetes complications have been well documented. For example, lifelong glycaemic control - a known risk factor for DKD - is not well captured in most existing cohorts and could not be used. Similarly, the use of antihypertensive over the duration of follow-up is not captured in all cohorts. Ideally, a time dependent analysis using time to event models should be used to investigate the genetic loci associated with DKD with adjustments for drugs and other relevant covariates. Further, the diabetic patient is exposed to many nonspecific kidney-damaging events in the course of disease (e.g. contrast agent imaging) which can cause transient renal dysfunction giving risk to albuminuria or reduced eGFR. Overall, the misclassification involved in using exclusively clinical DKD definitions in cross-sectional studies reduces statistical power to detect underlying genetic variants.

Additional Confounding factors

DKD is a complex disorder produced by the interplay of large number factors and additional confounding because of these can reduce the power of GWAS to detect novel associations. Long-term glycaemic control (measure of HbA1c) is the most

important factor, which can affect the onset of DKD. Given the cross-sectional nature of our study, we could not adjust for variations/fluctuations in HbA1c during the follow-up period. Similarly, we did not adjust for relevant covariates like insulin dose, diet, and exercise. Although the individual GWAS analysis across the SUMMIT cohorts was performed in homogenous population and were adjusted for eigenvectors from PCA, we cannot rule out confounding by population stratification.

Generalizability and Effect sizes

We performed the GWAS meta-analysis in European Caucasian populations and these results cannot be generalized to Asian, African and other admixed populations(African Americans, Hispanics) etc. Hence, these SNPs if replicated in Caucasians will have to be investigated in different ethnicities before they can be established as causal variants in all the diabetes population in the world. Although we have not estimated the effect sizes of these SNPs (as they are yet to be replicated), indications are that they will explain small proportion of variability of DKD in these populations. For, example the top SNP in *UMOD* explains less than 1% variation in eGFR in patients with diabetes. Despite of this very large GWAS effort to detect the genetic determinants of DKD, it is likely that we will be able to explain a small proportion of heritability of DKD.

Functional analysis

With this meta-analysis of DN in T1D patients and T2D DN patients, we have taken an important step towards defining the genetic architecture of DN. Strengths of the study include its large sample size, consideration of alternative DN phenotypes based albuminuria and eGFR staging. Now the challenge will be further investigations and

interpretations of these results in the context of pathophysiology of DKD. Since we have not replicated the GWAS hits for DKD, we have still not undertaken a functional analysis of the identified SNPs. How these SNPs interact in vivo to contribute to the pathogenesis of DKD can be studied using mouse models, tissue specific expressions once the SNPs are replicated in independent cohorts.

Future Directions

Genome-wide meta-analysis for Type 1 and Type 2 Diabetes in the SUMMIT consortium have revealed several potential loci associated with DKD. A logical next step would be to replicate these loci in independent cohorts. Given the huge amount of data accumulated due to meta-analysis of multiple phenotypes for DKD it is a challenge to select a right number of SNPs for replication. Top SNPs for replication can be selected in several ways. A straightforward way to do this would be to select the top independent SNPs for all phenotypes at a suggestive GWAS significance threshold ($P<10-6$) and carry them forward for replication. Other way would be to look at these top associations and selected SNPs that lie in biologically relevant pathways for DKD. Both these approaches have limitations. If only the top SNPs i.e. with a GWAS significance threshold ($P<10-6$) were selected, we would miss a number of important SNPs which are significant in multiple phenotypes at a significance of $P<10-5$. These could be relevant for DKD given that they are consistently significant across various stages or albuminuria. On the other-hand if we were to select SNPs based on their representation in biologically relevant candidates for DKD we will lose, a large number of loci –given that all the pathways for DKD have not been elucidated.

One way to deal with the multiple correlated phenotypes would be to use a new multivariate method that TATES (Trait-based Association Test that uses Extended Simes procedure), inspired by the GATES procedure proposed by Li et al (2011)[218]. For each component of a multivariate trait, TATES combines p-values obtained in standard univariate GWAS to acquire one trait-based p-value, while correcting for correlations between components. Extensive simulations, probing a wide variety of genotype–phenotype models, show that TATES's false positive rate is correct, and that TATES's statistical power to detect causal variants explaining 0.5% of the variance can be 2.5–9 times higher than the power of univariate tests based on composite scores and 1.5–2 times higher than the power of the standard MANOVA. Unlike other multivariate methods, TATES detects both genetic variants that are common to multiple phenotypes and genetic variants that are specific to a single phenotype, i.e. TATES provides a more complete view of the genetic architecture of complex traits. As the actual causal genotype–phenotype model is usually unknown and probably phenotypically and genetically complex, TATES, available as an open source program, constitutes a powerful new multivariate strategy that allows researchers to identify novel causal variants.

We would also need to perform a genome-wide conditional analysis on the summary level data accumulated through these GWAS analysis. It is likely that some SNP with big effect can mask SNPs with smaller effect sizes. For example, a GWAS for DKD stratified by effect of *MHY9* gene uncovered the association of *FRMD3* with DKD[114].

Perhaps the greatest challenge in the 'post-GWAS' era is to understand the functional consequences of the associated loci. Biological insights from functional analysis can then be translated to clinical benefits, screening, and disease prevention.

Our understanding of the way in which a risk variant initiates disease pathogenesis progresses from statistical association between genetic variation and trait or disease variation to functionality and causality. The functional consequences of variants in protein-coding regions causing most monogenic disorders are easy to interpret, as we know the genetic code. For non-Mendelian or multifactorial traits, most of the common DNA variants have so far mapped to non-protein–coding regions, where our understanding of functional consequences and causality is extremely limited. One way, that the trait-associated alleles exert their effects is by being a transcription activator (such as transcript levels and splicing) through multiple mechanisms. A transcriptional activator is a protein that increases gene transcription of a gene or set of gene. After confirmation of the associated loci, (following replication studies) we can use appropriate assays and models to test the functional effects of both SNPs and genes associated with DKD.

In summary, the future work will involve confirmation of the loci through replications in independent populations, studying the effect of the replicated loci in different ethnicities and ascertaining the mechanism through which these loci affect the pathogenesis of DKD.

In conclusion this thesis has shown that a) estimation of chip based heritability of various DKD sub-phenotypes using GCTA has limited utility and requires GWAS studies with extremely large sample sizes b) the genetic determinants of renal function (eGFR) can interact with albuminuria in patients with T2D c) there are yet unidentified genetic markers associated with DKD and have identified potentially novel genetic markers associated with sRAGE (an important biomarker for DKD) and DKD itself which can be investigated in future studies for their reproducibility and functional consequences.

References

1. Sherwin R, Jastreboff AM. Year in diabetes 2012: The diabetes tsunami. *J Clin Endocrinol Metab.* Dec 2012;97(12):4293-4301.
2. Mogensen CE, Chachati A, Christensen CK, et al. Microalbuminuria: an early marker of renal involvement in diabetes. *Uremia Invest.* 1985;9(2):85-95.
3. Mathiesen ER, Ronn B, Storm B, Foght H, Deckert T. The natural course of microalbuminuria in insulin-dependent diabetes: a 10-year prospective study. *Diabet Med.* Jun 1995;12(6):482-487.
4. Perkins BA, Ficociello LH, Silva KH, Finkelstein DM, Warram JH, Krolewski AS. Regression of microalbuminuria in type 1 diabetes. *N Engl J Med.* Jun 5 2003;348(23):2285-2293.
5. Adler AI, Stevens RJ, Manley SE, Bilous RW, Cull CA, Holman RR. Development and progression of nephropathy in type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS 64). *Kidney Int.* Jan 2003;63(1):225-232.
6. Inker LA, Astor BC, Fox CH, et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. *Am J Kidney Dis.* May 2014;63(5):713-735.
7. Krolewski AS, Niewczas MA, Skupien J, et al. Early progressive renal decline precedes the onset of microalbuminuria and its progression to macroalbuminuria. *Diabetes Care.* Jan 2014;37(1):226-234.
8. Pham TT, Sim JJ, Kujubu DA, Liu IL, Kumar VA. Prevalence of nondiabetic renal disease in diabetic patients. *Am J Nephrol.* 2007;27(3):322-328.
9. . *Chronic Kidney Disease: National Clinical Guideline for Early Identification and Management in Adults in Primary and Secondary Care*. London2008.
10. Fioretto P, Mauer M. Histopathology of diabetic nephropathy. *Semin Nephrol.* Mar 2007;27(2):195-207.
11. Brocco E, Fioretto P, Mauer M, et al. Renal structure and function in non-insulin dependent diabetic patients with microalbuminuria. *Kidney Int Suppl.* Dec 1997;63:S40-44.
12. Girach A, Vignati L. Diabetic microvascular complications--can the presence of one predict the development of another? *J Diabetes Complications.* Jul-Aug 2006;20(4):228-237.
13. Marshall SM. Recent advances in diabetic nephropathy. *Postgrad Med J.* Nov 2004;80(949):624-633.
14. Caramori ML, Fioretto P, Mauer M. The need for early predictors of diabetic nephropathy risk: is albumin excretion rate sufficient? *Diabetes.* Sep 2000;49(9):1399-1408.
15. MacLeod JM, Lutale J, Marshall SM. Albumin excretion and vascular deaths in NIDDM. *Diabetologia.* May 1995;38(5):610-616.
16. Young BA, Maynard C, Boyko EJ. Racial differences in diabetic nephropathy, cardiovascular disease, and mortality in a national population of veterans. *Diabetes Care.* Aug 2003;26(8):2392-2399.
17. Pavkov ME, Knowler WC, Hanson RL, Nelson RG. Diabetic nephropathy in American Indians, with a special emphasis on the Pima Indians. *Curr Diab Rep.* Dec 2008;8(6):486-493.
18. Dasmahapatra A, Bale A, Raghuwanshi MP, et al. Incipient and overt diabetic nephropathy in African Americans with NIDDM. *Diabetes Care.* Apr 1994;17(4):297-304.
19. Klein R, Klein BE, Moss SE, Davis MD, DeMets DL. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. IX. Four-year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch Ophthalmol.* Feb 1989;107(2):237-243.
20. Osterby R. Morphometric studies of the peripheral glomerular basement membrane in early juvenile diabetes. I. Development of initial basement membrane thickening. *Diabetologia.* Apr 1972;8(2):84-92.
21. Fioretto P, Steffes MW, Mauer M. Glomerular structure in nonproteinuric IDDM patients with various levels of albuminuria. *Diabetes.* Nov 1994;43(11):1358-1364.
22. Mauer SM, Steffes MW, Ellis EN, Sutherland DE, Brown DM, Goetz FC. Structural-functional relationships in diabetic nephropathy. *J Clin Invest.* Oct 1984;74(4):1143-1155.
23. Barisoni L, Mundel P. Podocyte biology and the emerging understanding of podocyte diseases. *Am J Nephrol.* Sep-Oct 2003;23(5):353-360.
24. Dahlquist G, Stattin EL, Rudberg S. Urinary albumin excretion rate and glomerular filtration rate in the prediction of diabetic nephropathy; a long-term follow-up study of childhood onset type-1 diabetic patients. *Nephrol Dial Transplant.* Jul 2001;16(7):1382-1386.
25. Melsom T, Mathisen UD, Ingebretsen OC, et al. Impaired fasting glucose is associated with renal hyperfiltration in the general population. *Diabetes Care.* Jul 2011;34(7):1546-1551.

26. Nelson RG, Bennett PH, Beck GJ, et al. Development and progression of renal disease in Pima Indians with non-insulin-dependent diabetes mellitus. Diabetic Renal Disease Study Group. *N Engl J Med.* Nov 28 1996;335(22):1636-1642.

27. Magee GM, Bilous RW, Cardwell CR, Hunter SJ, Kee F, Fogarty DG. Is hyperfiltration associated with the future risk of developing diabetic nephropathy? A meta-analysis. *Diabetologia.* Apr 2009;52(4):691-697.

28. Zatz R, Dunn BR, Meyer TW, Anderson S, Rennke HG, Brenner BM. Prevention of diabetic glomerulopathy by pharmacological amelioration of glomerular capillary hypertension. *J Clin Invest.* Jun 1986;77(6):1925-1930.

29. Cherney DZ, Scholey JW, Miller JA. Insights into the regulation of renal hemodynamic function in diabetic mellitus. *Curr Diabetes Rev.* Nov 2008;4(4):280-290.

30. Sharma K, Ziyadeh FN. Hyperglycemia and diabetic kidney disease. The case for transforming growth factor-beta as a key mediator. *Diabetes.* Oct 1995;44(10):1139-1146.

31. Aires I, Calado J. BI-10773, a sodium-glucose cotransporter 2 inhibitor for the potential oral treatment of type 2 diabetes mellitus. *Curr Opin Investig Drugs.* Oct 2010;11(10):1182-1190.

32. Sasson AN, Cherney DZ. Renal hyperfiltration related to diabetes mellitus and obesity in human disease. *World J Diabetes.* Jan 15 2012;3(1):1-6.

33. Arakawa K, Ishihara T, Oku A, et al. Improved diabetic syndrome in C57BL/KsJ-db/db mice by oral administration of the Na(+)-glucose cotransporter inhibitor T-1095. *Br J Pharmacol.* Jan 2001;132(2):578-586.

34. Strojek K, Yoon KH, Hruba V, Elze M, Langkilde AM, Parikh S. Effect of dapagliflozin in patients with type 2 diabetes who have inadequate glycaemic control with glimepiride: a randomized, 24-week, double-blind, placebo-controlled trial. *Diabetes Obes Metab.* Oct 2011;13(10):928-938.

35. Hohenstein B, Hausknecht B, Boehmer K, Riess R, Brekken RA, Hugo CP. Local VEGF activity but not VEGF expression is tightly regulated during diabetic nephropathy in man. *Kidney Int.* May 2006;69(9):1654-1661.

36. De Vriese AS, Tilton RG, Stephan CC, Lameire NH. Vascular endothelial growth factor is essential for hyperglycemia-induced structural and functional alterations of the peritoneal membrane. *J Am Soc Nephrol.* Aug 2001;12(8):1734-1741.

37. Isaka Y, Akagi Y, Ando Y, Imai E. Application of gene therapy to diabetic nephropathy. *Kidney Int Suppl.* Sep 1997;60:S100-103.

38. Dronavalli S, Duka I, Bakris GL. The pathogenesis of diabetic nephropathy. *Nat Clin Pract Endocrinol Metab.* Aug 2008;4(8):444-452.

39. Benigni A, Zoja C, Corna D, et al. Add-on anti-TGF-beta antibody to ACE inhibitor arrests progressive diabetic nephropathy in the rat. *J Am Soc Nephrol.* Jul 2003;14(7):1816-1824.

40. Sharma K, Eltayeb BO, McGowan TA, et al. Captopril-induced reduction of serum levels of transforming growth factor-beta1 correlates with long-term renoprotection in insulin-dependent diabetic patients. *Am J Kidney Dis.* Nov 1999;34(5):818-823.

41. Chen ZJ, Yang YB, Huang SM. [Expression of VEGF in kidney of diabetic rats]. *Sichuan Da Xue Xue Bao Yi Xue Ban.* Jul 2007;38(4):633-636.

42. Eremina V, Cui S, Gerber H, et al. Vascular endothelial growth factor a signaling in the podocyte-endothelial compartment is required for mesangial cell migration and survival. *J Am Soc Nephrol.* Mar 2006;17(3):724-735.

43. Navarro-Gonzalez JF, Mora-Fernandez C. The role of inflammatory cytokines in diabetic nephropathy. *J Am Soc Nephrol.* Mar 2008;19(3):433-442.

44. Wendt TM, Tanji N, Guo J, et al. RAGE drives the development of glomerulosclerosis and implicates podocyte activation in the pathogenesis of diabetic nephropathy. *Am J Pathol.* Apr 2003;162(4):1123-1137.

45. Yamagishi S, Matsui T, Nakamura K. Kinetics, role and therapeutic implications of endogenous soluble form of receptor for advanced glycation end products (sRAGE) in diabetes. *Curr Drug Targets.* Oct 2007;8(10):1138-1143.

46. Tanji N, Markowitz GS, Fu C, et al. Expression of advanced glycation end products and their cellular receptor RAGE in diabetic nephropathy and nondiabetic renal disease. *J Am Soc Nephrol.* Sep 2000;11(9):1656-1666.

47. Yamamoto Y, Kato I, Doi T, et al. Development and prevention of advanced diabetic nephropathy in RAGE-overexpressing mice. *J Clin Invest.* Jul 2001;108(2):261-268.

48. Hansen TK, Tarnow L, Thiel S, et al. Association between mannose-binding lectin and vascular complications in type 1 diabetes. *Diabetes.* Jun 2004;53(6):1570-1576.

49. Bierhaus A, Schiekofer S, Schwaninger M, et al. Diabetes-associated sustained activation of the transcription factor nuclear factor-kappaB. *Diabetes.* Dec 2001;50(12):2792-2808.

50. Vasavada N, Agarwal R. Role of oxidative stress in diabetic nephropathy. *Adv Chronic Kidney Dis.* Apr 2005;12(2):146-154.

51. Suzuki D, Miyata T, Nangaku M, et al. Expression of megsin mRNA, a novel mesangium-predominant gene, in the renal tissues of various glomerular diseases. *J Am Soc Nephrol.* Dec 1999;10(12):2606-2613.

52. Trevisan R, Dodesini AR, Lepore G. Lipids and renal disease. *J Am Soc Nephrol.* Apr 2006;17(4 Suppl 2):S145-147.

53. Vaziri ND. Molecular mechanisms of lipid disorders in nephrotic syndrome. *Kidney Int.* May 2003;63(5):1964-1976.

54. Abrass CK. Cellular lipid metabolism and the role of lipids in progressive renal disease. *Am J Nephrol.* Jan-Feb 2004;24(1):46-53.

55. Imig JD. Eicosanoids and renal vascular function in diseases. *Clin Sci (Lond).* Jul 2006;111(1):21-34.

56. Seaquist ER, Goetz FC, Rich S, Barbosa J. Familial clustering of diabetic kidney disease. Evidence for genetic susceptibility to diabetic nephropathy. *N Engl J Med.* May 4 1989;320(18):1161-1165.

57. Borch-Johnsen K, Norgaard K, Hommel E, et al. Is diabetic nephropathy an inherited complication? *Kidney Int.* Apr 1992;41(4):719-722.

58. Pettitt DJ, Saad MF, Bennett PH, Nelson RG, Knowler WC. Familial predisposition to renal disease in two generations of Pima Indians with type 2 (non-insulin-dependent) diabetes mellitus. *Diabetologia.* Jul 1990;33(7):438-443.

59. Viswanathan M, McCarthy MI, Snehalatha C, Hitman GA, Ramachandran A. Familial aggregation of type 2 (non-insulin-dependent) diabetes mellitus in south India; absence of excess maternal transmission. *Diabet Med.* Mar 1996;13(3):232-237.

60. Canani LH, Gerchman F, Gross JL. Increased familial history of arterial hypertension, coronary heart disease, and renal disease in Brazilian type 2 diabetic patients with diabetic nephropathy. *Diabetes Care.* Sep 1998;21(9):1545-1550.

61. Ramirez SP, McClellan W, Port FK, Hsu SI. Risk factors for proteinuria in a large, multiracial, southeast Asian population. *J Am Soc Nephrol.* Jul 2002;13(7):1907-1917.

62. Freedman BI, Volkova NV, Satko SG, et al. Population-based screening for family history of end-stage renal disease among incident dialysis patients. *Am J Nephrol.* Nov-Dec 2005;25(6):529-535.

63. Satko SG, Langefeld CD, Daeihagh P, Bowden DW, Rich SS, Freedman BI. Nephropathy in siblings of African Americans with overt type 2 diabetic nephropathy. *Am J Kidney Dis.* Sep 2002;40(3):489-494.

64. Basford KE, De Lacy IH. The use of matrix specifications in defining gene action in genotypic value models and generation mean analysis. *Theor Appl Genet.* Sep 1979;55(5):225-229.

65. Forsblom CM, Kanninen T, Lehtovirta M, Saloranta C, Groop LC. Heritability of albumin excretion rate in families of patients with Type II diabetes. *Diabetologia.* Nov 1999;42(11):1359-1366.

66. Fogarty DG, Rich SS, Hanna L, Warram JH, Krolewski AS. Urinary albumin excretion in families with type 2 diabetes is heritable and genetically correlated to blood pressure. *Kidney Int.* Jan 2000;57(1):250-257.

67. Langefeld CD, Beck SR, Bowden DW, Rich SS, Wagenknecht LE, Freedman BI. Heritability of GFR and albuminuria in Caucasians with type 2 diabetes mellitus. *Am J Kidney Dis.* May 2004;43(5):796-800.

68. Freedman BI, Beck SR, Rich SS, et al. A genome-wide scan for urinary albumin excretion in hypertensive families. *Hypertension.* Sep 2003;42(3):291-296.

69. Bochud M, Elston RC, Maillard M, et al. Heritability of renal function in hypertensive families of African descent in the Seychelles (Indian Ocean). *Kidney Int.* Jan 2005;67(1):61-69.

70. Placha G, Poznik GD, Dunn J, et al. A genome-wide linkage scan for genes controlling variation in renal function estimated by serum cystatin C levels in extended families with type 2 diabetes. *Diabetes.* Dec 2006;55(12):3358-3365.

71. Coresh J, Astor BC, Greene T, Eknoyan G, Levey AS. Prevalence of chronic kidney disease and decreased kidney function in the adult US population: Third National Health and Nutrition Examination Survey. *Am J Kidney Dis.* Jan 2003;41(1):1-12.

72. Ritz E, Zeng XX, Rychlik I. Clinical manifestation and natural history of diabetic nephropathy. *Contrib Nephrol.* 2011;170:19-27.

73. Breyer MD, Bottinger E, Brosius FC, 3rd, et al. Mouse models of diabetic nephropathy. *J Am Soc Nephrol.* Jan 2005;16(1):27-45.

74. Alpers CE, Hudkins KL. Mouse models of diabetic nephropathy. *Curr Opin Nephrol Hypertens.* May 2011;20(3):278-284.

75. Tesch GH, Lim AK. Recent insights into diabetic renal injury from the db/db mouse model of type 2 diabetic nephropathy. *Am J Physiol Renal Physiol.* Feb 2011;300(2):F301-310.

76. Crisan D, Carr J. Angiotensin I-converting enzyme: genotype and disease associations. *J Mol Diagn.* Aug 2000;2(3):105-115.

77. Marre M, Bernadet P, Gallois Y, et al. Relationships between angiotensin I converting enzyme gene polymorphism, plasma levels, and diabetic retinal and renal complications. *Diabetes.* Mar 1994;43(3):384-388.

78. Powrie JK, Watts GF, Ingham JN, Taub NA, Talmud PJ, Shaw KM. Role of glycaemic control in development of microalbuminuria in patients with insulin dependent diabetes. *Bmj.* Dec 17 1994;309(6969):1608-1612.
79. Tarnow L, Cambien F, Rossing P, et al. Lack of relationship between an insertion/deletion polymorphism in the angiotensin I-converting enzyme gene and diabetic nephropathy and proliferative retinopathy in IDDM patients. *Diabetes.* May 1995;44(5):489-494.
80. Schmidt S, Schone N, Ritz E. Association of ACE gene polymorphism and diabetic nephropathy? The Diabetic Nephropathy Study Group. *Kidney Int.* Apr 1995;47(4):1176-1181.
81. Mizuiri S, Hemmi H, Inoue A, et al. Angiotensin-converting enzyme polymorphism and development of diabetic nephropathy in non-insulin-dependent diabetes mellitus. *Nephron.* 1995;70(4):455-459.
82. Dudley CR, Keavney B, Stratton IM, Turner RC, Ratcliffe PJ. U.K. Prospective Diabetes Study. XV: Relationship of renin-angiotensin system gene polymorphisms with microalbuminuria in NIDDM. *Kidney Int.* Dec 1995;48(6):1907-1911.
83. Marre M, Jeunemaitre X, Gallois Y, et al. Contribution of genetic polymorphism in the renin-angiotensin system to the development of renal complications in insulin-dependent diabetes: Genetique de la Nephropathie Diabetique (GENEDIAB) study group. *J Clin Invest.* Apr 1 1997;99(7):1585-1595.
84. Suzuki H, Nagase S, Kikuchi S, Wang Y, Koyama A. Association of a missense Glu298Asp mutation of the endothelial nitric oxide synthase gene with end stage renal disease. *Clin Chem.* Nov 2000;46(11):1858-1860.
85. Nagase S, Suzuki H, Wang Y, et al. Association of ecNOS gene polymorphisms with end stage renal diseases. *Mol Cell Biochem.* Feb 2003;244(1-2):113-118.
86. Tolins JP, Palmer RM, Moncada S, Raij L. Role of endothelium-derived relaxing factor in regulation of renal hemodynamic responses. *Am J Physiol.* Mar 1990;258(3 Pt 2):H655-662.
87. He Y, Fan Z, Zhang J, et al. Polymorphisms of eNOS gene are associated with diabetic nephropathy: a meta-analysis. *Mutagenesis.* Mar 2011;26(2):339-349.
88. Tian GX, Zeng XT, Wang XB, Zhang L, Zhang W, Wei WL. Association between the endothelial nitric oxide synthase gene Glu298Asp polymorphism and coronary heart disease: a metaanalysis of 39 casecontrol studies. *Mol Med Rep.* Apr 2013;7(4):1310-1318.
89. Ilhan N, Kahraman N, Seckin D, Colak R. Apo E gene polymorphism on development of diabetic nephropathy. *Cell Biochem Funct.* Sep-Oct 2007;25(5):527-532.
90. Mooyaart AL, Valk EJ, van Es LA, et al. Genetic associations in diabetic nephropathy: a meta-analysis. *Diabetologia.* Mar 2011;54(3):544-553.
91. Ziyadeh FN. Different roles for TGF-beta and VEGF in the pathogenesis of the cardinal features of diabetic nephropathy. *Diabetes Res Clin Pract.* Nov 13 2008;82 Suppl 1:S38-41.
92. Kopp JB, Factor VM, Mozes M, et al. Transgenic mice with increased plasma levels of TGF-beta 1 develop progressive renal disease. *Lab Invest.* Jun 1996;74(6):991-1003.
93. Sharma K, Jin Y, Guo J, Ziyadeh FN. Neutralization of TGF-beta by anti-TGF-beta antibody attenuates kidney hypertrophy and the enhanced extracellular matrix gene expression in STZ-induced diabetic mice. *Diabetes.* Apr 1996;45(4):522-530.
94. Shaker OG, Sadik NA. Transforming growth factor beta 1 and monocyte chemoattractant protein-1 as prognostic markers of diabetic nephropathy. *Hum Exp Toxicol.* Oct 2013;32(10):1089-1096.
95. Prasad P, Tiwari AK, Kumar KM, et al. Association of TGFbeta1, TNFalpha, CCR2 and CCR5 gene polymorphisms in type-2 diabetes and renal insufficiency among Asian Indians. *BMC Med Genet.* 2007;8:20.
96. McKnight AJ, Savage DA, Patterson CC, Sadlier D, Maxwell AP. Resequencing of genes for transforming growth factor beta1 (TGFB1) type 1 and 2 receptors (TGFBR1, TGFBR2), and association analysis of variants with diabetic nephropathy. *BMC Med Genet.* 2007;8:5.
97. Sandholm N, Salem RM, McKnight AJ, et al. New susceptibility loci associated with kidney disease in type 1 diabetes. *PLoS Genet.* Sep 2012;8(9):e1002921.
98. He B, Osterholm AM, Hoverfalt A, et al. Association of genetic variants at 3q22 with nephropathy in patients with type 1 diabetes mellitus. *American journal of human genetics.* Jan 2009;84(1):5-13.
99. Igo RP, Jr., Iyengar SK, Nicholas SB, et al. Genomewide linkage scan for diabetic renal failure and albuminuria: the FIND study. *Am J Nephrol.* 2011;33(5):381-389.
100. Bowden DW, Colicigno CJ, Langefeld CD, et al. A genome scan for diabetic nephropathy in African Americans. *Kidney Int.* Oct 2004;66(4):1517-1526.
101. Vardarli I, Baier LJ, Hanson RL, et al. Gene for susceptibility to diabetic nephropathy in type 2 diabetes maps to 18q22.3-23. *Kidney Int.* Dec 2002;62(6):2176-2183.
102. McDonough CW, Bostrom MA, Lu L, et al. Genetic analysis of diabetic nephropathy on chromosome 18 in African Americans: linkage analysis and dense SNP mapping. *Hum Genet.* Dec 2009;126(6):805-817.

103. Imperatore G, Hanson RL, Pettitt DJ, Kobes S, Bennett PH, Knowler WC. Sib-pair linkage analysis for susceptibility genes for microvascular complications among Pima Indians with type 2 diabetes. Pima Diabetes Genes Group. *Diabetes.* May 1998;47(5):821-830.
104. Krolewski AS, Poznik GD, Placha G, et al. A genome-wide linkage scan for genes controlling variation in urinary albumin excretion in type II diabetes. *Kidney Int.* Jan 2006;69(1):129-136.
105. Iyengar SK, Abboud HE, Goddard KA, et al. Genome-wide scans for diabetic nephropathy and albuminuria in multiethnic populations: the family investigation of nephropathy and diabetes (FIND). *Diabetes.* Jun 2007;56(6):1577-1585.
106. Chen G, Adeyemo A, Zhou J, et al. Genome-wide search for susceptibility genes to type 2 diabetes in West Africans: potential role of C-peptide. *Diabetes Res Clin Pract.* Dec 2007;78(3):e1-6.
107. Freedman BI, Bowden DW, Rich SS, et al. Genome-wide linkage scans for renal function and albuminuria in Type 2 diabetes mellitus: the Diabetes Heart Study. *Diabet Med.* Mar 2008;25(3):268-276.
108. Moczulski DK, Rogus JJ, Antonellis A, Warram JH, Krolewski AS. Major susceptibility locus for nephropathy in type 1 diabetes on chromosome 3q: results of novel discordant sib-pair analysis. *Diabetes.* Jul 1998;47(7):1164-1169.
109. Osterholm AM, He B, Pitkaniemi J, et al. Genome-wide scan for type 1 diabetic nephropathy in the Finnish population reveals suggestive linkage to a single locus on chromosome 3q. *Kidney Int.* Jan 2007;71(2):140-145.
110. Rogus JJ, Poznik GD, Pezzolesi MG, et al. High-density single nucleotide polymorphism genome-wide linkage scan for susceptibility genes for diabetic nephropathy in type 1 diabetes: discordant sibpair approach. *Diabetes.* Sep 2008;57(9):2519-2526.
111. Iyengar SK, Elston RC. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods Mol Biol.* 2007;376:71-84.
112. Pezzolesi MG, Poznik GD, Mychaleckyj JC, et al. Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. *Diabetes.* Jun 2009;58(6):1403-1410.
113. Nathan DM, Bayless M, Cleary P, et al. Diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: advances and contributions. *Diabetes.* Dec 2013;62(12):3976-3986.
114. McDonough CW, Palmer ND, Hicks PJ, et al. A genome-wide association study for diabetic nephropathy genes in African Americans. *Kidney Int.* Mar 2011;79(5):563-572.
115. Freedman BI, Langefeld CD, Lu L, et al. Differential effects of MYH9 and APOL1 risk variants on FRMD3 Association with Diabetic ESRD in African Americans. *PLoS Genet.* Jun 2011;7(6):e1002150.
116. Kao WH, Klag MJ, Meoni LA, et al. MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nature genetics.* Oct 2008;40(10):1185-1192.
117. Maeda S, Osawa N, Hayashi T, Tsukada S, Kobayashi M, Kikkawa R. Genetic variations associated with diabetic nephropathy and type II diabetes in a Japanese population. *Kidney Int Suppl.* Aug 2007(106):S43-48.
118. Hanson RL, Craig DW, Millis MP, et al. Identification of PVT1 as a candidate gene for end-stage renal disease in type 2 diabetes using a pooling-based genome-wide single nucleotide polymorphism association study. *Diabetes.* Apr 2007;56(4):975-983.
119. Craig DW, Millis MP, DiStefano JK. Genome-wide SNP genotyping study using pooled DNA to identify candidate markers mediating susceptibility to end-stage renal disease attributed to Type 1 diabetes. *Diabet Med.* Nov 2009;26(11):1090-1098.
120. Remuzzi G, Benigni A, Remuzzi A. Mechanisms of progression and regression of renal lesions of chronic nephropathies and diabetes. *J Clin Invest.* Feb 2006;116(2):288-296.
121. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics.* Jan 7 2011;88(1):76-82.
122. Pearson ER, Donnelly LA, Kimber C, et al. Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study. *Diabetes.* Aug 2007;56(8):2178-2182.
123. Morris AD, Boyle DI, McMahon AD, Greene SA, MacDonald TM, Newton RW. Adherence to insulin treatment, glycaemic control, and ketoacidosis in insulin-dependent diabetes mellitus. The DARTS/MEMO Collaboration. Diabetes Audit and Research in Tayside Scotland. Medicines Monitoring Unit. *Lancet.* Nov 22 1997;350(9090):1505-1510.
124. Chaturvedi N, Stephenson JM, Fuller JH. The relationship between socioeconomic status and diabetes control and complications in the EURODIAB IDDM Complications Study. *Diabetes Care.* May 1996;19(5):423-430.
125. Colhoun HM, Thomason MJ, Mackness MI, et al. Design of the Collaborative AtoRvastatin Diabetes Study (CARDS) in patients with type 2 diabetes. *Diabet Med.* Mar 2002;19(3):201-211.
126. Morris AD, Boyle DI, MacAlpine R, et al. The diabetes audit and research in Tayside Scotland (DARTS) study: electronic record linkage to create a diabetes register. DARTS/MEMO Collaboration. *Bmj.* Aug 30 1997;315(7107):524-528.

127. Doney AS, Dannfald J, Kimber CH, et al. The FTO gene is associated with an atherogenic lipid profile and myocardial infarction in patients with type 2 diabetes: a Genetics of Diabetes Audit and Research Study in Tayside Scotland (Go-DARTS) study. *Circ Cardiovasc Genet.* Jun 2009;2(3):255-259.

128. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann Intern Med.* Mar 16 1999;130(6):461-470.

129. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med.* May 5 2009;150(9):604-612.

130. Michels WM, Grootendorst DC, Verduijn M, Elliott EG, Dekker FW, Krediet RT. Performance of the Cockcroft-Gault, MDRD, and new CKD-EPI formulas in relation to GFR, age, and body size. *Clin J Am Soc Nephrol.* Jun 2010;5(6):1003-1009.

131. Kukla A, Issa N, Jackson S, et al. Cystatin C enhances glomerular filtration rate estimating equations in kidney transplant recipients. *Am J Nephrol.* 2014;39(1):59-65.

132. Zhou K, Bellenguez C, Spencer CC, et al. Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. *Nature genetics.* Feb 2011;43(2):117-120.

133. Tesfaye S, Chaturvedi N, Eaton SE, et al. Vascular risk factors and diabetic neuropathy. *N Engl J Med.* Jan 27 2005;352(4):341-350.

134. Microvascular and acute complications in IDDM patients: the EURODIAB IDDM Complications Study. *Diabetologia.* Mar 1994;37(3):278-285.

135. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* Jun 2009;5(6):e1000529.

136. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nature methods.* Feb 2012;9(2):179-181.

137. Southam L, Panoutsopoulou K, Rayner NW, et al. The effect of genome-wide association scan quality control on imputation outcome for common variants. *Eur J Hum Genet.* May 2011;19(5):610-614.

138. Conrad DF, Keebler JE, DePristo MA, et al. Variation in genome-wide mutation rates within and between human families. *Nature genetics.* Jul 2011;43(7):712-714.

139. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet.* Oct 2011;12(10):703-714.

140. Delaneau O, Coulonges C, Zagury JF. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics.* 2008;9:540.

141. Patterson N, Hattangadi N, Lane B, et al. Methods for high-density admixture mapping of disease genes. *American journal of human genetics.* May 2004;74(5):979-1000.

142. Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol.* Apr 2008;32(3):215-226.

143. Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* Dec 1999;55(4):997-1004.

144. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *American journal of human genetics.* Jul 2000;67(1):170-181.

145. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet.* Oct 15 2008;17(R2):R122-128.

146. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* Jan 7 2011;88(1):76-82.

147. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* Jul 2010;42(7):565-569.

148. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics.* Apr 2012;44(4):369-375, S361-363.

149. Feng S, Wang S, Chen CC, Lan L. GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits. *BMC Genet.* 2011;12:12.

150. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* Sep 1 2010;26(17):2190-2191.

151. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics.* Sep 2007;81(3):559-575.

152. Zhang L, Krzentowski G, Albert A, Lefebvre PJ. Factors predictive of nephropathy in DCCT Type 1 diabetic patients with good or poor metabolic control. *Diabet Med.* Jul 2003;20(7):580-585.

153. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics.* Feb 2006;38(2):209-213.

154. Yang J, Lee SH, Goddard ME, Visscher PM. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol Biol.* 2013;1019:215-236.

155. Chang A, Van Horn L, Jacobs DR, Jr., et al. Lifestyle-related factors, obesity, and incident microalbuminuria: the CARDIA (Coronary Artery Risk Development in Young Adults) study. *Am J Kidney Dis.* Aug 2013;62(2):267-275.
156. Taal MW, Brenner BM. Predicting initiation and progression of chronic kidney disease: Developing renal risk scores. *Kidney Int.* Nov 2006;70(10):1694-1705.
157. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. *N Engl J Med.* Sep 30 1993;329(14):977-986.
158. Stratton IM, Adler AI, Neil HA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *Bmj.* Aug 12 2000;321(7258):405-412.
159. Shulman NB, Ford CE, Hall WD, et al. Prognostic value of serum creatinine and effect of treatment of hypertension on renal function. Results from the hypertension detection and follow-up program. The Hypertension Detection and Follow-up Program Cooperative Group. *Hypertension.* May 1989;13(5 Suppl):I80-93.
160. Retnakaran R, Cull CA, Thorne KI, Adler AI, Holman RR. Risk factors for renal dysfunction in type 2 diabetes: U.K. Prospective Diabetes Study 74. *Diabetes.* Jun 2006;55(6):1832-1839.
161. Gray R, Wheatley K. How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant.* 1991;7 Suppl 3:9-12.
162. Salvi E, Kutalik Z, Glorioso N, et al. Genomewide association study using a high-density single nucleotide polymorphism array and case-control design identifies a novel essential hypertension susceptibility locus in the promoter region of endothelial NO synthase. *Hypertension.* Feb 2012;59(2):248-255.
163. Newton-Cheh C, Johnson T, Gateva V, et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nature genetics.* Jun 2009;41(6):666-676.
164. Padmanabhan S, Melander O, Johnson T, et al. Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS Genet.* Oct 2010;6(10):e1001177.
165. Levy D, Ehret GB, Rice K, et al. Genome-wide association study of blood pressure and hypertension. *Nature genetics.* Jun 2009;41(6):677-687.
166. Soranzo N, Sanna S, Wheeler E, et al. Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. *Diabetes.* Dec 2010;59(12):3229-3239.
167. Pare G, Chasman DI, Parker AN, et al. Novel association of HK1 with glycated hemoglobin in a non-diabetic population: a genome-wide evaluation of 14,618 participants in the Women's Genome Health Study. *PLoS Genet.* Dec 2008;4(12):e1000312.
168. Paterson AD, Waggott D, Boright AP, et al. A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose. *Diabetes.* Feb 2010;59(2):539-549.
169. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* Jan 2014;42(Database issue):D1001-1006.
170. Reznichenko A, Boger CA, Snieder H, et al. UMOD as a susceptibility gene for end-stage renal disease. *BMC Med Genet.* 2012;13(1):78.
171. Kottgen A, Yang Q, Shimmin LC, et al. Association of estimated glomerular filtration rate and urinary uromodulin concentrations with rare variants identified by UMOD gene region sequencing. *PLoS One.* 2012;7(5):e38311.
172. Ahluwalia TS, Lindholm E, Groop L, Melander O. Uromodulin gene variant is associated with type 2 diabetic nephropathy. *J Hypertens.* Sep 2011;29(9):1731-1734.
173. Gudbjartsson DF, Holm H, Indridason OS, et al. Association of variants at UMOD with chronic kidney disease and kidney stones-role of age and comorbid diseases. *PLoS Genet.* Jul 2010;6(7):e1001039.
174. Boger CA, Gorski M, Li M, et al. Association of eGFR-Related Loci Identified by GWAS with Incident CKD and ESRD. *PLoS Genet.* Sep 2011;7(9):e1002292.
175. Liu CT, Garnaas MK, Tin A, et al. Genetic association for renal traits among participants of African ancestry reveals new loci for renal function. *PLoS Genet.* Sep 2011;7(9):e1002264.
176. Bonetti S, Trombetta M, Boselli ML, et al. Variants of GCKR affect both beta-cell and kidney function in patients with newly diagnosed type 2 diabetes: the Verona newly diagnosed type 2 diabetes study 2. *Diabetes Care.* May 2011;34(5):1205-1210.
177. Shriner D, Herbert A, Doumatey AP, et al. Multiple loci associated with renal function in African Americans. *PLoS One.* 2012;7(9):e45112.
178. Kottgen A, Glazer NL, Dehghan A, et al. Multiple loci associated with indices of renal function and chronic kidney disease. *Nature genetics.* Jun 2009;41(6):712-717.
179. Kottgen A, Pattaro C, Boger CA, et al. New loci associated with kidney function and chronic kidney disease. *Nature genetics.* May 2010;42(5):376-384.

180. Gaens KH, Ferreira I, van der Kallen CJ, et al. Association of polymorphism in the receptor for advanced glycation end products (RAGE) gene with circulating RAGE levels. *J Clin Endocrinol Metab.* Dec 2009;94(12):5174-5180.
181. Jang Y, Kim JY, Kang SM, et al. Association of the Gly82Ser polymorphism in the receptor for advanced glycation end products (RAGE) gene with circulating levels of soluble RAGE and inflammatory markers in nondiabetic and nonobese Koreans. *Metabolism.* Feb 2007;56(2):199-205.
182. Raucci A, Cugusi S, Antonelli A, et al. A soluble form of the receptor for advanced glycation endproducts (RAGE) is produced by proteolytic cleavage of the membrane-bound form by the sheddase a disintegrin and metalloprotease 10 (ADAM10). *Faseb J.* Oct 2008;22(10):3716-3727.
183. Zhang L, Bukulin M, Kojro E, et al. Receptor for advanced glycation end products is subjected to protein ectodomain shedding by metalloproteinases. *J Biol Chem.* Dec 19 2008;283(51):35507-35516.
184. Li K, Dai D, Zhao B, et al. Association between the RAGE G82S polymorphism and Alzheimer's disease. *J Neural Transm.* Jan 2010;117(1):97-104.
185. Forbes JM, Soderlund J, Yap FY, et al. Receptor for advanced glycation end-products (RAGE) provides a link between genetic susceptibility and environmental factors in type 1 diabetes. *Diabetologia.* May 2011;54(5):1032-1042.
186. Hirota T, Takahashi A, Kubo M, et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. *Nature genetics.* Sep 2011;43(9):893-896.
187. Salonen KM, Ryhanen SJ, Forbes JM, et al. Circulating Concentrations of Soluble Receptor for AGE Are Associated With Age and AGER Gene Polymorphisms in Children With Newly Diagnosed Type 1 Diabetes. *Diabetes Care.* Apr 17 2014.
188. Lasky-Su J, Himes BE, Raby BA, et al. HLA-DQ strikes again: genome-wide association study further confirms HLA-DQ in the diagnosis of asthma among adults. *Clin Exp Allergy.* Dec 2012;42(12):1724-1733.
189. Smith DJ, Yerkovich ST, Towers MA, Carroll ML, Thomas R, Upham JW. Reduced soluble receptor for advanced glycation end-products in COPD. *Eur Respir J.* Mar 2011;37(3):516-522.
190. Sukkar MB, Wood LG, Tooze M, et al. Soluble RAGE is deficient in neutrophilic asthma and COPD. *Eur Respir J.* Mar 2012;39(3):721-729.
191. Katakami N, Matsuhisa M, Kaneto H, et al. Decreased endogenous secretory advanced glycation end product receptor in type 1 diabetic patients: its possible association with diabetic vascular complications. *Diabetes Care.* Nov 2005;28(11):2716-2721.
192. Lee HJ, Kim SY, Koh JM, et al. Polymorphisms and haplotypes of integrinalpha1 (ITGA1) are associated with bone mineral density and fracture risk in postmenopausal Koreans. *Bone.* Dec 2007;41(6):979-986.
193. Chambers JC, Zhang W, Sehmi J, et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature genetics.* Nov 2011;43(11):1131-1138.
194. Billings LK, Hsu YH, Ackerman RJ, et al. Impact of common variation in bone-related genes on type 2 diabetes and related traits. *Diabetes.* Aug 2012;61(8):2176-2186.
195. ITGA1 gene locus may shed light on osteoporosis-type 2 diabetes link. *Bonekey Rep.* 2012;1:248.
196. Sparvero LJ, Asafu-Adjei D, Kang R, et al. RAGE (Receptor for Advanced Glycation Endproducts), RAGE ligands, and their role in cancer and inflammation. *J Transl Med.* 2009;7:17.
197. Pettersson-Fernholm K, Forsblom C, Hudson BI, Perola M, Grant PJ, Groop PH. The functional -374 T/A RAGE gene polymorphism is associated with proteinuria and cardiovascular disease in type 1 diabetic patients. *Diabetes.* Mar 2003;52(3):891-894.
198. Falcone C, Campo I, Emanuele E, et al. Relationship between the -374T/A RAGE gene polymorphism and angiographic coronary artery disease. *Int J Mol Med.* Dec 2004;14(6):1061-1064.
199. dos Santos KG, Canani LH, Gross JL, Tschiedel B, Pires Souto KE, Roisenberg I. The -374A allele of the receptor for advanced glycation end products gene is associated with a decreased risk of ischemic heart disease in African-Brazilians with type 2 diabetes. *Mol Genet Metab.* Jun 2005;85(2):149-156.
200. Wang J, Zou L, Song Z, et al. Meta-analysis of RAGE gene polymorphism and coronary heart disease risk. *PLoS One.* 2012;7(12):e50790.
201. Laki J, Kiszel P, Vatay A, et al. The HLA 8.1 ancestral haplotype is strongly linked to the C allele of -429T>C promoter polymorphism of receptor of the advanced glycation endproduct (RAGE) gene. Haplotype-independent association of the -429C allele with high hemoglobinA1C levels in diabetic patients. *Mol Immunol.* Jan 2007;44(4):648-655.
202. Sullivan CM, Futers TS, Barrett JH, Hudson BI, Freeman MS, Grant PJ. RAGE polymorphisms and the heritability of insulin resistance: the Leeds family study. *Diab Vasc Dis Res.* Feb 2005;2(1):42-44.

203. Looker HC, Nyangoma SO, Cromie DT, et al. Predicted impact of extending the screening interval for diabetic retinopathy: the Scottish Diabetic Retinopathy Screening programme. *Diabetologia.* Aug 2013;56(8):1716-1725.
204. Kwon JH, Kim YH, Kwon JY, Park YW. Clinical significance of serum sRAGE and esRAGE in women with normal pregnancy and preeclampsia. *J Perinat Med.* Sep 2011;39(5):507-513.
205. Okada Y, Kamatani Y, Takahashi A, et al. Common variations in PSMD3-CSF3 and PLCB4 are associated with neutrophil count. *Hum Mol Genet.* May 15 2010;19(10):2079-2085.
206. Reddan DN, Klassen PS, Szczech LA, et al. White blood cells as a novel mortality predictor in haemodialysis patients. *Nephrol Dial Transplant.* Jun 2003;18(6):1167-1173.
207. Green EK, Grozeva D, Moskvina V, et al. Variation at the GABAA receptor gene, Rho 1 (GABRR1) associated with susceptibility to bipolar schizoaffective disorder. *Am J Med Genet B Neuropsychiatr Genet.* Oct 5 2010;153B(7):1347-1349.
208. Xuei X, Flury-Wetherill L, Dick D, et al. GABRR1 and GABRR2, encoding the GABA-A receptor subunits rho1 and rho2, are associated with alcohol dependence. *Am J Med Genet B Neuropsychiatr Genet.* Mar 5 2010;153B(2):418-427.
209. Lin HJ, Huang YC, Lin JM, et al. Novel susceptibility genes associated with diabetic cataract in a Taiwanese population. *Ophthalmic Genet.* Mar-Jun 2013;34(1-2):35-42.
210. Chung SS, Ho EC, Lam KS, Chung SK. Contribution of polyol pathway to diabetes-induced oxidative stress. *J Am Soc Nephrol.* Aug 2003;14(8 Suppl 3):S233-236.
211. Zhang KW, Tonna S, Wang YY, Rana K, Padavarat S, Savige J. Do mutations in COL4A1 or COL4A2 cause thin basement membrane nephropathy (TBMN)? *Pediatr Nephrol.* May 2007;22(5):645-651.
212. Schrijvers BF, Flyvbjerg A, De Vriese AS. The role of vascular endothelial growth factor (VEGF) in renal pathophysiology. *Kidney Int.* Jun 2004;65(6):2003-2017.
213. Vasir B, Jonas JC, Steil GM, et al. Gene expression of VEGF and its receptors Flk-1/KDR and Flt-1 in cultured and transplanted rat islets. *Transplantation.* Apr 15 2001;71(7):924-935.
214. Veron D, Bertuccio CA, Marlier A, et al. Podocyte vascular endothelial growth factor (Vegf(1)(6)(4)) overexpression causes severe nodular glomerulosclerosis in a mouse model of type 1 diabetes. *Diabetologia.* May 2011;54(5):1227-1241.
215. Spittau B, Krieglstein K. Klf10 and Klf11 as mediators of TGF-beta superfamily signaling. *Cell Tissue Res.* Jan 2012;347(1):65-72.
216. Barton A, Eyre S, Ke X, et al. Identification of AF4/FMR2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further pan-autoimmune susceptibility genes. *Hum Mol Genet.* Jul 1 2009;18(13):2518-2522.
217. Brown D, Wagner CA. Molecular mechanisms of acid-base sensing by the kidney. *J Am Soc Nephrol.* May 2012;23(5):774-780.
218. van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet.* 2013;9(1):e1003235.

Appendix I

Script for running IMPUTE 2 and GTOOL

```
lineval=`awk 'match($1,/^[0-9]/){

interval=5000;

count++;

   if (count==1)  {

      start=$1;

   }

   if (count%(interval)==0 && count >1) {

      end=$1;

      system("qsub             -cwd -q 64bit.q -b y          -l ram=8000M ./impute2 -m
/homes/hcolhoun/Eurodiab/shapeit/genetic_map_chr11_combined_b37.txt  -known_haps_g
/homes/hcolhoun/Eurodiab/shapeit/chr11.haps   -h
/homes/hcolhoun/Eurodiab/shapeit/ALL_1000G_phase1integrated_v3_chr11_impute.hap.gz   -l
/homes/hcolhoun/Eurodiab/shapeit/ALL_1000G_phase1integrated_v3_chr11_impute.legend.gz  -Ne
20000 -int " start "  " end " -o /homes/hcolhoun/Eurodiab/shapeit/chr11/OUTPUT_FILE_chunk_"
start "  -allow_large_regions  -seed 367946");

      print $cmdline;

   }

   if ((count-1)%(interval)==0) {

      start=$1;

   }

}' $1`


echo $lineval;
```

Script for running QC on SNPTEST results in R

```
data<-
read.table('hld.adjustedbaseline_test',header=T,stringsAsFactors=FALSE,sep=
" ")
data1<-
subset(data,select=c("X10","rsid","pos","allele_A","allele_B","average_maxi
mum_posterior_call","all_AA","all_AB","all_BB","all_maf","cohort_1_hwe","De
ltaLn_frequentist_add_xage_xsex_LogB_score_pvalue","DeltaLn_frequentist_add
_xage_xsex_LogB_score_info","DeltaLn_frequentist_add_xage_xsex_LogB_score_b
eta_1","DeltaLn_frequentist_add_xage_xsex_LogB_score_se_1"))
data2<-
data1[data1$all_maf>0.01&data1$DeltaLn_frequentist_add_xage_xsex_LogB_score
_info>0.60,]
data2$cohort_1_hwe<-as.numeric(as.character(data2$cohort_1_hwe))
data3<-
data2[!(data2$all_maf<0.05&data2$cohort_1_hwe<0.0001)&!(data2$all_maf>0.05&
data2$cohort_1_hwe<0.00000057),]
```

```
names(data3)<-
c("CHR","MARKER","POS","NEA","EA","CALL_RATE","N0","N1","N2","MAF","HWE_P",
"PVAL","INFO","BETA","SE")
data3$BETA<-as.numeric(data3$BETA)
data3$SE<-as.numeric(data3$SE)
data3$INFO<-as.numeric(data3$INFO)
data3$N0<-as.numeric(data3$N0)
data3$N1<-as.numeric(data3$N1)
data3$N2<-as.numeric(data3$N2)
EAF=(2*data3$N2+data3$N1)/(2*data3$N0+2*data3$N1+2*data3$N2)
N=data3$N0+data3$N1+data3$N2
data4<-
subset(data3,select=c("id","MARKER","POS","NEA","EA","HWE_P","PVAL","INFO",
"BETA","SE","CALL_RATE"))
data5<-cbind(data4,EAF,N)
data8<-data5[!(data5$PVAL<0) & !(data5$INFO<0),]
data9<-
subset(data8,select=c("id","MARKER","POS","NEA","EA","EAF","N","HWE_P","PVA
L","INFO","BETA","SE","CALL_RATE"))
write.table(data9,file='hld.adjustedbaseline_out_qc',sep="
",row.names=FALSE,quote=FALSE)
data8$PVAL<-as.numeric(data8$PVAL)
Nsummary<-summary(data8$N)
Psummary<-summary(data8$PVAL)
infosummary<-summary(data8$INFO)
BETAsummary<-summary(data8$BETA)
totalsummary<-rbind(Nsummary,Psummary,infosummary,BETAsummary)
write.table(totalsummary,file="hld.adjustedbaseline.summary",sep="
",row.names=FALSE,quote=FALSE)
```

Investigation of known estimated glomerular filtration rate loci in patients with Type 2 diabetes

DOI: 10.1111/dme.12211

Short Report: Genetics

Investigation of known estimated glomerular filtration rate loci in patients with Type 2 diabetes

H. A. Deshmukh[1], C. N. A. Palmer[2], A. D. Morris[2] and H. M. Colhoun[1]

[1]Division of Population Health Sciences and [2]Division of Cardiovascular and Diabetes Medicine, University of Dundee, Dundee, UK

Accepted 10 April 2013

Abstract

Aims To replicate the association of genetic variants with estimated glomerular filtration rate (GFR) and albuminuria, which has been found in recent genome-wide studies in patients with Type 2 diabetes.

Methods We evaluated 16 candidate single nucleotide polymorphisms for estimated GFR in 3028 patients with Type 2 diabetes sampled from clinics across Tayside, Scotland, UK, who were included in the Genetics of Diabetes Audit and Research Tayside (GoDARTs) study. These single nucleotide polymorphisms were tested for their association with estimated GFR at entry to the study, with albuminuria, and with time to stage 3B chronic kidney disease (estimated GFR<45 ml/min/1.73 m^2). We also stratified the effects on estimated GFR in patients with (n = 2096) and without albuminuria (n = 613).

Results rs1260326 in *GCKR* (β=1.30, P = 3.23E-03), rs17319721 in *SHROOM3* (β = −1.28, P-value = 3.18E-03) and rs12917707 in *UMOD* (β = 2.0, P-value = 8.84E-04) were significantly associated with baseline estimated GFR. Analysis of effects on estimated GFR, stratified by albuminuria status, showed that in those without albuminuria (normoalbuminura; n = 613), *UMOD* had a significantly stronger effect on estimated GFR (β_{normo} = 4.03 ± 1.23 vs $\beta_{albuminuria}$ = 1.72 ± 0.76, P = 0.002) compared with those with albuminuria, while *GCKR* (β_{normo} = 0.45 ± 0.89 vs $\beta_{albuminuria}$ = 1.12 ± 0.55, P = 0.08) and *SHROOM3* (β_{normo} = −0.07 ± 0.89 vs $\beta_{albuminuria}$ = −1.43 ± 0.53, P = 0.003) had a stronger effect on estimated GFR in those with albuminuria. *UMOD* was also associated with a lower rate of transition to stage 3B chronic kidney disease (hazard ratio = 0.83[0.70, 0.99], P = 0.03).

Conclusion The genetic variants that regulate estimated GFR in the general population tend to have similar effects in patients with Type 2 diabetes and in this latter population, it is important to adjust for albuminuria status while investigating the genetic determinants of renal function.

Diabet. Med. 30, 1230–1235 (2013)

Introduction

Recent genome-wide association studies have identified several genetic variants associated with estimated (e)GFR and chronic kidney disease (CKD). Previous investigations of these eGFR polymorphisms were typically carried out in populations where < 10% of patients were diagnosed with Type 2 diabetes [1]. It remains to be established if these variants are associated with eGFR in patients with Type 2 diabetes for whom there are different reasons for loss of renal function, in particular diabetic nephropathy, when compared with patients without diabetes. Most of these studies are cross-sectional [2–5], and so clinically relevant dynamic phenotypes cannot be studied. Longitudinal datasets capturing renal function can be used to investigate if the genetic variants identified are associated with a rapid decline in renal function (end-stage renal disease or stage 3 CKD) in patients with Type 2 diabetes. About 20% of patients with Type 2 diabetes with CKD defined according to the ADA guidelines may have normoalbuminuria (albumin/creatinine ratio [ACR] <2.5 mg/mmol in males and ACR<3.5 mg/mmol in females) [6]. The genetic and pathological mechanisms that determine the relationship between reduced eGFR and albuminuria status in patients with Type 2 diabetes remain unknown [7]. Although the genetic variants associated with eGFR do not seem to be associated with albuminuria [8], it remains to be seen if these genetic variants have the same effect on eGFR in

Correspondence to: Harshal A. Deshmukh. E-mail: h.deshmukh@dundee.ac.uk

What's new?

- This is the first study comparing common genetic variants associated with estimated GFR between the general population and patients with Type 2 diabetes.
- This is the first report of the interaction of genetic effects of estimated GFR-associated loci (UMOD GCKR and SHROOM3) with albuminuria in patients with Type 2 diabetes.
- The study stresses the need to adjust for albuminuria while investigating the genetic determinants of renal function.

those with and without albuminuria. In the present study, using a longitudinal cohort of patients with Type 2 diabetes, we investigated the association of 16 recently identified eGFR-associated loci (*LASS2, GCKR, NAT8, TFDP2, SHROOM3, DAB2, SLC34A1, VEGFA, PRKAG2, ADAM28, PIP5K1B, ATXN2, DACH1, UBE2Q2, UMOD, SLC7A9*) with baseline eGFR albuminuria, and time to stage 3B CKD (eGFR<45 ml/min/1.73 m^2), in patients with Type 2 diabetes.

Methods

The study population comprised 3028 patients with Type 2 diabetes identified from an on-going study, the Genetics of Diabetes Audit and Research Tayside (GoDARTs) study, and recruited in Tayside, Scotland, UK, between 1 October 1997 and 1 March 2010. The baseline clinical characteristics of the GoDARTs subset included in the present analyses were very similar to the baseline clinical characteristics of the remaining GoDARTS cohort, except that those not included were slightly older and had a lower eGFR (Table 1); therefore, the subset of patients used for the present analysis was very representative of the entire GoDARTs cohort. Calculations for eGFR were made using the Modification of Diet in Renal Disease formula [9] which requires age, sex, race and creatinine data. We assessed the association of the 16 single nucleotide polymorphisms (SNPs) with eGFR at baseline by linear regression analysis using the gPLINK program [10], adjusting for age, sex, BMI, population structure, HbA_{1c}, duration of diabetes and systolic blood pressure. To investigate whether the association of these loci with eGFR differed according to albuminuria status, we carried out a stratified analysis in patients with sustained normoalbuminuria (ACR <2.5 mg/mmol in males and <3.5 mg/mmol in females and with a duration of diabetes >15 years at end of follow-up) and in those with any albuminuria (ACR $\geq$2.5 mg/mmol in males and $\geq$3.5 mg/mmol in females, either at baseline or at the end of follow-up).

To investigate if any of these SNPs were associated with a rapid decline in renal function over the follow-up period, we performed an analysis of time to stage 3B CKD (eGFR<45 ml/min/1.73 m^2). Individuals with stage 3B CKD at baseline were excluded. By using this threshold, 4% of our patients were excluded from the analysis. If we had chosen to study progression to stage 3A CKD (eGFR<60 ml/min/1.73 m^2), 20% of patients would have been excluded from the analysis. Stage 3B CKD was defined as three consecutive eGFR measurements of eGFR <45 ml/min/1.73 m^2 at least 1 month apart. Those who did not progress to stage 3B CKD were censored at the end of the follow-up period or at date of death. We used a Cox proportional hazards model (the Proc PHREG tool in the SAS statistical package), with date of birth as 'time in' and 'last date' as the first date of eGFR <45 ml/min/1.73 m^2 or the end of follow-up period/date of death, and with genotype, age, sex, BMI and baseline eGFR as covariates. The interaction of individual SNPs with albuminuria was tested using PLINK option 'interaction' with age, sex, BMI, albuminuria and genotypes as covariates in the linear regression model. We adopted a conservative threshold for significance (0.05/number of loci tested) and a P value < 0.003 was considered to indicate statistical significance. A weighted genetic risk score analysis was performed to test the joint effect of the 16 loci on baseline eGFR and time to stage 3B CKD. We calculated weighted genetic risk score (number of risk alleles*β) for each individual using all 16 SNPs, and tested the association of this genetic risk score with baseline eGFR and time to stage 3B CKD, adjusting for age, sex, BMI, HbA_{1c}, duration of diabetes, and systolic blood pressure. All analyses were performed in PLINK version 1.07 [10] and SAS

Table 1 Demographic characteristics of the GoDARTs cohort

Characteristic, mean (SD)	GoDARTs cohort in the present study	GoDARTs cohort not included in the present study
Age at baseline, years	59.1 (11.0)	66.2 (11.6)
Sex,% female	46.4	42.3
Baseline BMI	30.6 (5.3)	31.5 (6.1)
Baseline eGFR, ml/min/1.73m^2	73.9 (18.7)	70.9 (15.8)
Baseline systolic blood pressure, mmHg	142.8 (18.4)	141.7 (18.8)
Baseline HbA_{1C}, mmol/mol	7.54 (1.3) (58 mmol/mol)*	7.3 (1.4) (56 mmol/mol)*
Baseline cholesterol, mmol/L	4.40 (0.97)	4.34 (0.91)
Duration of diabetes at baseline, years	8.71 (7.44)	7.75 (6.61)

*These are HbA1c values in IFCC units.

9.2. Power calculations for quantitative traits were performed using R 2.15.

Samples were genotyped at Affymetrix's service laboratory on the Genome-Wide Human SNP Array 6.0. Complete genotype data have been described previously [11]. The study complied with the Declaration of Helsinki guidelines. Since October 1997, all individuals with diabetes in the GoDARTs database have been invited to give consent for DNA collection as part of the Wellcome Trust United Kingdom Type 2 Diabetes case–control collection. As of June 2009, 8000 cases and 7000 control subjects of European ancestry have participated in this GoDARTS study. Informed consent was obtained from all the study participants.

Results

Table 1 shows the baseline characteristics of the GoDARTs cohort included in the present study as well as the GoDARTs cohort not genotyped at the conception of this study. Genotype data were available for 3028 patients (46.4% females) with Type 2 diabetes. Their mean (SD) baseline BMI was 30.6 (5.3) kg/m^2, mean (SD) age was 59.1 (11) years, mean (SD) HbA_{1c} was 58 mmol/mol (7.54 (±1.3). The mean (SD) follow-up period for the entire study was 10.6 (9.1) years with a median of three eGFR readings/year/person (interquartile range 2–4) and a mean (SD) baseline eGFR of 73.9 (18.7) ml/min/1.73 m^2.

Table 2 shows the association found for the 16 eGFR-associated loci with baseline eGFR and albuminuria; the study population was stratified by albuminuria status and the association of these SNPs with time to stage 3B CKD. The minor alleles 'T' of *GCKR* (β = 1.30, *P*-value = 3.23E-03), and 'T' of *UMOD* (β = 2.0 *P*-value = 8.84E-04) were associated with a higher eGFR at baseline and the minor 'A' of *SHROOM3* (β = −1.28, *P*-value = 3.18E-03) was associated with a lower eGFR at the predefined threshold ($P \leq 0.003$). None of the other SNPs was associated with baseline eGFR. None of the 16 SNPs included in the study were associated with albuminuria after correction for multiple testing (data not shown). In patients with sustained normoalbuminuria ($n = 613$), minor allele 'T' of *UMOD* was associated with eGFR (β = 4.03, *P*-value = 1.10E-03), while in patients with albuminuria ($n = 2096$) minor allele 'T' of *GCKR* (β = 1.12, *P*-value = 4.27E-02) and 'A' of *SHROOM3* (β = −1.43, *P*-value = 7.28E-03) were associated with eGFR. Of the 16 SNPs, *UMOD* (hazard ratio = 0.83(0.70, 0.99), *P*-value = 0.03), *PIP5K1B* (hazard ratio = 0.85(0.75, 0.96), *P*-value = 0.01) and *SLC7A9* (hazard ratio = 0.86(0.76, 0.98) *P*-value = 0.02) was associated with time to stage 3B CKD (eGFR<45 mls/min/1.73 m2) at the 0.05 threshold for significance. Although the *PIP5K1B* locus was significant at $P < 0.05$, the direction of effect was not consistent with a previous report by Köttgen *et al.* [4] and hence this cannot be regarded as a positive replication of this SNP for its association with eGFR and time to CKD stage 3B.

Since the variants tested in this study are associated with age-related decline in eGFR in general population (and not with any disease-specific decline) we used time-to-event analysis with date of birth as the starting point; however, we performed a sensitivity analysis in which we used the baseline of GoDARTs study as the starting point. Although this analysis decreases power because of a reduction in the person-years follow-up, we see a similar effect size of association with progression to stage 3B CKD. For example, the hazard ratio of *UMOD* with time to stage 3B CKD with the starting point as the GoDARTs study baseline (hazard ratio = 0.87(0.74, 1.03) *P*-value = 0.1) is very similar to the hazard ratio with date of birth as a starting point. The weighted genetic risk score for the 16 SNPS explained the 1% variation in baseline eGFR and was significantly associated with baseline eGFR after adjustments for age, sex, BMI, HbA_{1c}, duration of diabetes and systolic blood pressure ($P = 0.0026$, β = 0.84(±0.28). The weighted genetic risk score was not associated with time to stage 3B CKD ($P = 0.52$).

Discussion

In the present study, we replicated the association of *UMOD, GCKR* and *SHROOM3* with eGFR in patients with Type 2 diabetes. The study confirms the findings of previous studies showing the association of *UMOD* with eGFR and diabetic nephropathy [12–15] and the association of *GCKR* and *SHROOM3* with eGFR [1,16,17]. A study by Gudbjartsson *et al.* [12] demonstrated the interaction of *UMOD* with age [15]; while another study could not replicate this interaction. In the present study, we did not observe an interaction of *UMOD* with age in patients with Type 2 diabetes (*P*-value = 0.84).

None of the other variants were associated with eGFR after correction for multiple testing; however, the direction of effect was consistent with the previous studies for all the statistically significant loci (*GCKR, SHROOM3, UMOD*) and for the loci that did not pass the threshold of significance (except *TFDP2* and *PIP5K1B*). Our study had limited power to estimate the effect of these variants on eGFR. Taken together, all these variants explain the 1.4–14% heritability of eGFR [5] (with each SNP contributing typically < 0.5% heritability of eGFR). Our study had 97% power to detect an association with a SNP explaining 0.5% variability in eGFR and anything below 0.5% can remain undetected. It is also possible that some of these SNPs are not the causal SNPs and because of varying linkage disequilibrium, structure in our population could not be detected. It is also possible that the effects of some of these SNPs were attenuated by diabetes or diabetic kidney disease and therefore were not associated with eGFR in this study.

Table 2 Association of the known single nucleotide polymorphisms with baseline estimated GFR, estimated GFR stratified by albuminuria status and time to stage 3B chronic kidney disease

CHR	Gene	SNP	Effect allele	Association with baseline eGFR (n = 2970)		Association with eGFR in patients with sustained normoalbuminuria† (n = 613)		Association with eGFR in patients with albuminuria (n = 2097)		Interaction term Heterogeneity *P*-value	Association with time to Stage 3B CKD (eGFR<45)*‡		Direction of effect in GoDARTs consistent with Köttgen *et al.* [4]
				β (SE)	*P*-value	β (SE)	*P*-value	β (SE)	*P*-value		Hazard ratio (CI)	*P*-value	
1	LASS2	rs267734	C	0.77 (±0.51)	1.30E-01	2.24 (±1.07)	3.63E-02	0.71 (±0.62)	2.57E-01	9.60E-02	1.12 (0.98,1.29)	7.00E-02	Yes
2	GCKR	rs1260326	T	1.30 (±0.44)	3.23E-03	0.45 (±0.89)	6.12E-01	1.12 (±0.55)	4.27E-02	8.70E-02	0.98 (0.86,1.11)	7.60E-01	Yes
2	NAT8	rs13538	G	0.40 (±0.51)	4.32E-01	0.55 (±1.12)	6.24E-01	0.29 (±0.62)	6.34E-01	8.92E-01	1.02 (1.023,1.027)	2.70E-01	Yes
3	TFDP2	rs347685	C	−0.51 (±0.48)	2.82E-01	0.54 (±0.97)	5.77E-01	−1.07 (±0.59)	6.76E-02	3.95E-01	0.96 (0.83,1.10)	5.50E-01	No
4	SHROOM3	rs17319721	A	−1.28 (±0.43)	3.18E-03	−0.07 (±0.89)	9.34E-01	−1.43 (±0.53)	7.28E-03	3.00E-03	1.02 (0.90,1.15)	6.90E-01	Yes
5	DAB2	rs11959928	A	−0.43 (±0.45)	3.39E-01	−1.45 (±0.90)	1.07E-01	−0.29 (±0.55)	5.99E-01	3.41E-01	0.97 (0.86,1.10)	7.00E-01	Yes
5	SLC34A1	rs6420094	G	−1.35 (±0.61)	2.74E-02	−2.92 (±1.24)	1.87E-02	−0.69 (±0.75)	3.60E-01	2.79E-01	0.93 (0.78.1.10)	4.00E-01	Yes
6	VEGFA	rs881858	G	0.54 (±0.48)	2.63E-01	1.31 (±1.01)	1.92E-01	1.34 (±0.59)	2.21E-02	4.40E-02	0.95 (0.83,1.08)	4.70E-01	Yes
7	PRKAG2	rs7805747	A	−0.31 (±0.49)	5.24E-01	−0.72 (±0.98)	4.62E-01	0.31 (±0.60)	6.02E-01	9.30E-01	1.03 (0.90.1.19)	6.00E-01	Yes
8	ADAM28	rs10109414	T	−0.51 (±0.44)	2.41E-01	−1.57 (±0.90)	8.17E-02	−0.17 (±0.54)	7.49E-01	5.10E-01	0.99 (0.87,1,12)	8.70E-01	Yes
9	PIP5K1B	rs4744712	A	0.09 (±0.44)	8.47E-01	1.71 (±0.91)	6.25E-02	−0.33 (±0.55)	5.41E-01	9.31E-01	0.85 (0.75,0.96)	1.00E-02	No
12	ATXN2	rs653178	T	0.20 (±0.42)	6.28E-01	0.71 (±0.85)	4.05E-01	−0.13 (±0.52)	8.09E-01	9.47E-01	0.95 (0.83,1.08)	9.50E-01	Yes
13	DACH1	rs626277	C	0.75 (±0.44)	9.14E-02	0.85 (±0.90)	3.46E-01	0.28 (±0.54)	6.02E-01	3.93E-01	0.98 (0.87,1.10)	7.50E-01	Yes
15	UBE2Q2	rs1394125	A	−0.86 (±0.53)	1.03E-01	−1.14 (±1.07)	2.89E-01	−0.86 (±0.65)	1.85E-01	2.68E-01	1.11 (0.96,1.28)	1.50E-01	Yes
16	UMOD	rs12917707	T	2.0 (±0.60)	8.84E-04	4.03 (±1.23)	1.10E-03	1.72 (±0.76)	2.30E-02	2.00E-03	0.83 (0.70.0.99)	3.00E-02	Yes
19	SLC7A9	rs12460876	C	0.24 (±0.51)	6.90E-01	0.58 (±0.94)	5.30E-01	0.29 (±0.57)	6.00E-01	4.50E-01	0.86 (0.76,0.98)	2.00E-02	Yes

*Adjusted for age at baseline, duration of diabetes, baseline-estimated GFR, systolic blood pressure, mean HbA_{1c} and mean BMI.
†Patients with normoalbumiuria at baseline and at the end of follow-up with a duration of diabetes >15 years.
‡Stage 3B CKD defined as three consecutive readings of eGFR <45 ml/min/1.73 m^2. Those already at stage 3B CKD at baseline were excluded for this analysis.
SNP, single nucleotide polymorphism; CKD, chronic kidney disease; CHR, chromosome.

We examined the association of the 16 loci with a decline in renal function using a Cox proportional hazard model and estimated the effect of these loci on time to stage 3B CKD (eGFR<45ml/min/1.73 m^2). Given the high mortality associated with diabetic nephropathy, cross-sectional studies are prone to survival bias, as patients with severe forms of nephropathy are less likely to be included. Hence, it is important to investigate the eGFR loci in a time-dependent manner. Of the 16 SNPs, none were associated with time to stage 3B CKD at the predefined threshold of 0.003, however, *UMOD* and *SLC7A9* were associated with time to stage 3B CKD at the threshold of 0.05 (with the direction of effects consistent with that reported previously). *UMOD* and *SLC7A9* have a stronger effect on baseline eGFR as compared with other markers suggesting that SNPs with a strong effect on baseline eGFR influence the decline in renal function over time.

We performed a stratified analysis to examine the effect of albuminuria on the known genetic associations with eGFR. In Type 2 diabetic, nephropathy, albuminuria may be more closely associated with decline in renal function and the impact of genetic determinants of eGFR may differ depending on the presence or absence of nephropathy; therefore, we examined the effects on eGFR stratified by albuminuria. There is a clear difference in the effect sizes in those with sustained normalbuminuria and those with albuminuria. For example, the *UMOD* has twice the effect in patients with sustained normalbuminuria as compared with those with albuminuria (P-interaction = 0.002) while *SHROOM3* (P-interaction = 0.003) and *GCKR* (P-interaction = 0.08) had larger effect sizes in those with albuminuria. It is known that kidney diseases characterized by albuminuria, such as diabetic nephropathy can have ultrafiltration and high eGFR in the early stage of disease, while those characterized by reduced renal function such as hypertensive kidney disease, may be manifested with normoalbumiuria because of the reduced renal efficiency [18,19]. Hence, studying the genetic determinants of eGFR without adjusting for albuminuria status or studying genetic determinants of albuminuria without accounting for eGFR can reduce the power of these studies to identify the true genetic effects. Cumulatively, eGFR-associated loci explain only a small fraction of the total heritable contribution eGFR and stratifying by albuminuria status in our existing genome-wide association study datasets [3–5] can help us to uncover the missing heritability. It is important to point out, however, that the interaction of albuminuria with the genetic variants associated with eGFR in patients with Type 2 diabetes seen in the present study is the first report of this interaction in patients with Type 2 diabetes and needs to be confirmed in an independent sample.

In summary, our results show that some of the genetic determinants of eGFR in the general population are common to patients with Type 2 diabetes; however, in patients with Type 2 diabetes it is essential to adjust for albuminuria status while investigating the genetic determinants of renal function.

Funding sources

The Wellcome Trust provides support for the Wellcome Trust United Kingdom Type 2 Diabetes Case Control Collection (the Go-DARTS study) and the Chief Scientist Office provides informatics support. The Wellcome Trust funds the Scottish Health Informatics Programme. The Wellcome Trust (084726/Z/08/Z, 085475/Z/08/Z, and 085475/B/08/Z) funded genome-wide genotyping as part of WTCCC2. The IMI SUMMIT programme supported HD.

Competing interests

None declared.

Acknowledgements

We are grateful to all the participants who took part in this study, to the general practitioners, and the Scottish School of Primary Care for their assistance in participant recruitment, and to the whole team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The study complies with the ethical guidelines as set out in the Declaration of Helsinki. We acknowledge the support of the Health Informatics Centre, University of Dundee for managing and supplying the anonymized data and NHS Tayside, the original data owner. We thank Miss Tahira Akbar for her help with editing the manuscript.

References

1 Böger CA, Gorski M, Li M, Hoffmann MM, Huang C, Yang Q *et al.* Association of eGFR-Related Loci Identified by GWAS with Incident CKD and ESRD. *PLoS Genet* 2011; **7**: e1002292.
2 O'Seaghdha CM, Fox CS. Genome-wide association studies of chronic kidney disease: what have we learned? *Nat Rev Nephrol* 2012; **8**: 89–99.
3 Köttgen A, Glazer NL, Dehghan A, Hwang SJ, Katz R, Li M *et al.* Multiple loci associated with indices of renal function and chronic kidney disease. *Nat Genet* 2009; **41**: 712–717.
4 Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL *et al.* New loci associated with kidney function and chronic kidney disease. *Nat Genet* 2010; **42**: 376–384.
5 Shriner D, Herbert A, Doumatey AP, Zhou J, Huang H, Erdos MR *et al.* Multiple loci associated with renal function in African Americans. *PloS One* 2012; **7**: e45112.
6 MacIsaac RJ, Tsalamandris C, Panagiotopoulos S, Smith TJ, McNeil KJ, Jerums G. Nonalbuminuric renal insufficiency in type 2 diabetes. *Diabetes Care* 2004; **27**: 195–200.
7 MacIsaac RJ, Panagiotopoulos S, McNeil KJ, Smith TJ, Tsalamandris C, Hao H. Is nonalbuminuric renal insufficiency in type 2 diabetes related to an increase in intrarenal vascular disease? *Diabetes Care* 2006; **29**: 1560–1566.

8 Ellis JW, Chen MH, Foster MC, Liu CT, Larson MG, de Boer I *et al*. Validated SNPs for eGFR and their associations with albuminuria. *Hum Mol Genet* 2012; **21**: 3293–3298.
9 Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann Intern Med* 1999; **130**: 461–70.
10 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
11 Bellenguez C, Strange A, Freeman C, Wellcome Trust Case Control Consortium, Donnelly P, Spencer CC. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* 2012; **28**: 134–135.
12 Reznichenko A, Böger CA, Snieder H, van den Born J, de Borst MH, Damman J *et al*. UMOD as a susceptibility gene for end-stage renal disease. *BMC Med Genet* 2012; **13**: 78.
13 Köttgen A, Yang Q, Shimmin LC, Tin A, Schaeffer C, Coresh J *et al*. Association of estimated glomerular filtration rate and urinary uromodulin concentrations with rare variants identified by UMOD gene region sequencing. *PloS One* 2012; **7**: e38311.
14 Ahluwalia TS, Lindholm E, Groop L, Melander O. Uromodulin gene variant is associated with type 2 diabetic nephropathy. *J Hypertens* 2011; **29**: 1731–1734.
15 Gudbjartsson DF, Holm H, Indridason OS, Thorleifsson G, Edvardsson V, Sulem P *et al*. Association of variants at UMOD with chronic kidney disease and kidney stones-role of age and comorbid diseases. *PLoS Genet* 2010; **6**: e1001039.
16 Liu CT, Garnaas MK, Tin A, Kottgen A, Franceschini N, Peralta CA *et al*. Genetic association for renal traits among participants of African ancestry reveals new loci for renal function. *PLoS Genet* 2011; **7**: e1002264.
17 Bonetti S, Trombetta M, Boselli ML, Turrini F, Malerba G, Trabetti E. Variants of GCKR affect both beta-cell and kidney function in patients with newly diagnosed type 2 diabetes: the Verona newly diagnosed type 2 diabetes study 2. *Diabetes Care* 2011; **34**: 1205–1210.
18 Coresh J, Astor BC, Greene T, Eknoyan G, Levey AS. Prevalence of chronic kidney disease and decreased kidney function in the adult US population: Third National Health and Nutrition Examination Survey. *Am J Kidney Dis* 2003; **41**: 1–12.
19 Ritz E, Zeng XX, Rychlík I. Clinical manifestation and natural history of diabetic nephropathy. *Contrib Nephrol* 2011; **170**: 19–27.

Genome-wide association study of genetic determinants of LDL-c response to atorvastatin therapy: importance of Lp(a)[S]

Harshal A. Deshmukh,[1,*] Helen M. Colhoun,[1,2,*] Toby Johnson,[†] Paul M. McKeigue,[§] D. John Betteridge,[**] Paul N. Durrington,[††] John H. Fuller,[**] Shona Livingstone,[*] Valentine Charlton-Menys,[††] Andrew Neil,[§§] Neil Poulter,[***] Peter Sever,[***] Denis C. Shields,[†††] Alice V. Stanton,[§§§] Aurobindo Chatterjee,[****] Craig Hyde,[****] Roberto A. Calle,[****] David A. DeMicco,[****] Stella Trompet,[††††] Iris Postmus,[§§§§] Ian Ford,[*****] J. Wouter Jukema,[††††,†††††] Mark Caulfield,[3,†] and Graham A. Hitman,[3,†] on behalf of the CARDS, ASCOT, and PROSPER investigators

University of Dundee,* Dundee, United Kingdom; Barts and The London School of Medicine and Dentistry,[†] Queen Mary University of London, London, United Kingdom; University of Edinburgh,[§] Edinburgh, United Kingdom; University College London,** London, United Kingdom; University of Manchester,[††] Manchester, United Kingdom; University of Oxford,[§§] Oxford, United Kingdom; International Centre for Circulatory Health,*** Imperial College London, United Kingdom; Complex and Adaptive Systems Laboratory,[†††] University College Dublin, Dublin, Ireland; Royal College of Surgeons in Ireland,[§§§] Dublin, Ireland; Pfizer Ltd.,**** New York, NY; Department of Cardiology[††††] and Department of Geriatrics and Gerontology,[§§§§] Leiden University Medical Center, Leiden, The Netherlands; Robertson Centre for Biostatistics,***** University of Glasgow, Glasgow, United Kingdom; and Interuniversity Cardiology Institute of the Netherlands,[†††††] Utrecht, The Netherlands

Abstract We carried out a genome-wide association study (GWAS) of LDL-c response to statin using data from participants in the Collaborative Atorvastatin Diabetes Study (CARDS; n = 1,156), the Anglo-Scandinavian Cardiac Outcomes Trial (ASCOT; n = 895), and the observational phase of ASCOT (n = 651), all of whom were prescribed atorvastatin 10 mg. Following genome-wide imputation, we combined data from the three studies in a meta-analysis. We found associations of LDL-c response to atorvastatin that reached genome-wide significance at rs10455872 ($P = 6.13 \times 10^{-9}$) within the *LPA* gene and at two single nucleotide polymorphisms (SNP) within the *APOE* region (rs445925; $P = 2.22 \times 10^{-16}$ and rs4420638; $P = 1.01 \times 10^{-11}$) that are proxies for the ε2 and ε4 variants, respectively, in APOE. The novel association with the *LPA* SNP was replicated in the PROspective Study of Pravastatin in the Elderly at Risk (PROSPER) trial ($P = 0.009$). Using CARDS data, we further showed that atorvastatin therapy did not alter lipoprotein(a) [Lp(a)] and that Lp(a) levels accounted for all of the associations of SNPs in the LPA gene and the apparent LDL-c response levels. However, statin therapy had a similar effect in reducing cardiovascular disease (CVD) in patients in the top quartile for serum Lp(a) levels (HR = 0.60) compared with those in the lower three quartiles (HR = 0.66; $P = 0.8$ for interaction). The data emphasize that high Lp(a) levels affect the measurement of LDL-c and the clinical estimation of LDL-c response. Therefore, an apparently lower LDL-c response to statin therapy may indicate a need for measurement of Lp(a). However, statin therapy seems beneficial even in those with high Lp(a).—Deshmukh, H. A., H. M. Colhoun, T. Johnson, P. M. McKeigue, D. J. Betteridge, P. N.

The CARDS trial was cofunded by Pfizer Ltd., Diabetes UK, and National Health Service R&D. Genotyping was funded by Pfizer Ltd. The Anglo-Scandinavian Cardiac Outcomes Trial and establishment of the associated genetic repository were funded by Pfizer Inc. Genotyping was funded by Barts, the London School of Medicine and Dentistry, and by the Centre Nationale de Genotypage Paris. The PROSPER study was supported by an investigator-initiated grant obtained from Bristol-Myers Squibb. Prof. Dr. J. W. Jukema is an established clinical investigator of the Netherlands Heart Foundation (Grant 2001-D-032). Support for genotyping was provided by the Seventh Framework Program of the European Commission (Grant 223004) and by the Netherlands Genomics Initiative (Netherlands Consortium for Healthy Aging Grant 050-060-810). For the PROSPER/PHASE program, the research leading to results received funding from the European Commission's Seventh Framework Program (FP7/2007–2013) under Grant HEALTH-F2-2009-223004 PHASE. CARDS registration number: NCT00327418. ASCOT registration number: EUDRACT2008-007494-20.

✤ ***Author's Choice**—Final version full access.*

Manuscript received 1 October 2011 and in revised form 24 January 2012.

Published, JLR Papers in Press, February 24, 2012
DOI 10.1194/jlr.P021113

Abbreviations: ASCOT, Anglo-Scandinavian Cardiac Outcomes Trial; CARDS, Collaborative Atorvastatin Diabetes Study; CVD, cardiovascular disease; GWAS, genome-wide association study; HWE, Hardy-Weinberg equilibrium; IQR, interquartile range; LD, linkage disequilibrium; Lp(a), lipoprotein(a); PROSPER, PROspective Study of Pravastatin in the Elderly at Risk; SNP, single nucleotide polymorphism; TC, total cholesterol.

[1] H. A. Deshmukh and H. M. Colhoun contributed equally to this work.
[2] To whom correspondence should be addressed.
e-mail: h.colhoun@cpse.dundee.ac.uk
[3] M. Caulfield and G. A. Hitman contributed equally to this work.

[S] The online version of this article (available at http://www.jlr.org) contains supplementary data in the form of three tables.

Durrington, J. H. Fuller, S. Livingstone, V. Charlton-Menys, A. Neil, N. Poulter, P. Sever, D. C. Shields, A. V. Stanton, A. Chatterjee, C. Hyde, R. A. Calle, D. A. DeMicco, S. Trompet, I. Postmus, I. Ford, J. W. Jukema, M. Caulfield, and G. A. Hitman on behalf of the CARDS, ASCOT, and PROSPER investigators. **Genome-wide association study of genetic determinants of LDL-c response to atorvastatin therapy: importance of Lp(a).** ***J. Lipid Res.*** **2012.** 53: **1000–1011.**

Supplementary key words genetics • low density lipoprotein • LDL/metabolism • lipoprotein(a) • statins

Statin therapy is now widely accepted for the primary and secondary prevention of cardiovascular disease (CVD) in certain patient groups. However, there is considerable variation in response to statin therapy that remains poorly understood. For example, in the Collaborative Atorvastatin Diabetes Study (CARDS) trial (1), among self-reported and pill count-validated compliant recipients of atorvastatin 10 mg daily, the absolute change in LDL-c at one month post-randomization varied from −2 to −0.6 mmol/l, (5^{th} and 95^{th} centiles of the range), and the percentage lowering from baseline varied from 67% to 22%. Understanding the pathways and determinants involved in this variation in response to therapy could lead to improved treatments. Even without understanding the pathways, identifying predictors of poorer response could identify those most in need of additional or alternative therapeutic strategies.

Two genome-wide association studies (GWAS) of statin response and several candidate gene association studies have been reported (2–5). From these, the only consistent finding is that variants in the *APOE* gene region are associated with variation in LDL response. Here, we report a genome-wide analysis of LDL-c response from two randomized clinical trials of atorvastatin, CARDS and the Anglo-Scandinavian Outcomes Trial (ASCOT) (6), to investigate genetic effects on LDL-c response to atorvastatin. We chose to model genetic determinants of LDL-c response to atorvastatin among those assigned to atorvastatin in these trials. An alternative approach would be to model the interaction of genotype on the effect of atorvastatin on LDL-c using data from both placebo and active treatment groups. However, we did not consider this latter approach as optimal as testing for interactions is much less powerful than direct tests of association and as, in any case, we did not consider genetic effects on change LDL-c in the placebo groups to be plausible.

MATERIALS AND METHODS

Study populations and phenotype definition

Both trials were conducted with Ethics Committee/IRB approval, under good clinical practice guidelines and in accordance with the Declaration of Helsinki principles. Patients gave consent for genetic studies.

CARDS

Methods in CARDS have been described previously. In brief, 2,838 patients with type 2 diabetes and no previous CVD were randomized to receive either placebo or atorvastatin 10 mg once daily and followed for a median of 3.7 years. Allocation was double blinded. Mean serum LDL-c concentration during baseline visits prior to randomization had to be ⩽ 4.14 mmol/l (160 mg/dl) and serum triglycerides ⩽ 6.78 mmol/l (600 mg/dl). After randomization, total cholesterol (TC), HDL-C, and triglycerides were measured at one, two, and three months, and then every six months. Patients attended after an overnight fast. LDL-c was calculated with the Friedewald formula (7), or if serum triglycerides exceeded 4.0 mmol/l, by removing VLDL by ultracentrifugation and then measuring the change in infranatant cholesterol content when LDL was removed by precipitation of apolipoprotein B-containing lipoproteins. For this genome-wide study, the analyses were restricted to those randomized to atorvastatin, and the mean of two pretreatment LDL-c measurements was used as the baseline LDL-c and a weighted average of five post-randomization values within the first year post-randomization was the outcome measure or "on treatment LDL-c," with weights (0.6 for month 1 and then 0.1 for measurements at 2, 3, 6, and 12 months). Lipoprotein(a) concentrations were determined by an immunoturbidimetric assay with Immuno LEIA® reagents from Technoclone Ltd., Dorking, UK (now www.PathwayDiagnostics.com), which is calibrated against the IFCC standard preparation PRM02.

ASCOT

Of 19,342 hypertensive patients (40–79 years of age with at least three other cardiovascular risk factors) who were randomized to one of two antihypertensive regimens in ASCOT, 10,305 with nonfasting TC concentrations of 6.5 mmol/l or less (measured at the nonfasting screening visit) had been randomly assigned additional atorvastatin 10 mg or placebo. These patients formed the lipid-lowering arm of the study. For this genome-wide study, two subpopulations from ASCOT were included. The first subpopulation included individuals randomized to 10 mg atorvastatin in whom pretreatment LDL-c was measured at the (fasting) randomization visit and on-treatment LDL-c was calculated as the simple average of measures at the (fasting) visits 6 months and 12 months post-randomization. LDL-c was estimated using the Friedewald equation as in CARDS. Following the end of the randomization phase, there was an observational period. The second subpopulation included all individuals not originally randomized to 10 mg atorvastatin (i.e., those randomized to placebo and those not eligible for the LLA) who were subsequently prescribed atorvastatin 10 mg. For these individuals, pretreatment LDL-c was defined as the measurement on the last visit before or equal to date of starting atorvastatin, and on-treatment LDL-c was defined as the measurement taken from the first visit after date of starting atorvastatin.

PROSPER (replication cohort)

All data were from the PROspective Study of Pravastatin in the Elderly at Risk (PROSPER) (8). PROSPER was a prospective multicenter randomized placebo-controlled trial to assess whether treatment with pravastatin diminishes the risk of major vascular events in elderly. Between December 1997 and May 1999, we screened and enrolled subjects in Scotland (Glasgow), Ireland (Cork), and the Netherlands (Leiden). Men and women 70–82 years of age were recruited if they had preexisting vascular disease or increased risk of such disease because of smoking, hypertension, or diabetes. A total number of 5,804 subjects were randomly assigned to pravastatin or placebo, of which 2,550 subjects assigned to the Pravastatin arm of the trial were included in the present study. TC, HDL-C, and triglycerides were assessed after an overnight fast, at baseline, and at 3, 6, 12, 24, and 36 months post-randomization. LDL-C was calculated by the Friedewald formula. The pretreatment measurement was at baseline before

randomization, and the posttreatment was the mean of the lipid measurements after randomization.

Phenotype transformation

To maximize power to detect associations and to improve test statistic behavior under the null for low minor allele frequency (MAF) single nucleotide polymorphisms (SNP), we transformed measured LDL-c levels to conform to the distributional assumptions made by our association analysis model using the same transformation for off- and on-treatment measures to preserve the relationship between the two. We maximized the fit of the residuals in a regression of on-treatment on the pretreatment value to a Gaussian distribution. We used a 2-parameter Box-Cox transform of the form $(x-\beta)^{\alpha}/\alpha$ applied to baseline and on-treatment LDL-c values. The parameter values α and β were chosen by maximizing the likelihood of a model with linear regressions of the transformed pretreatment and response (transformed pretreatment minus transformed pretreatment) values on the covariates (age and sex), with the joint distribution of the residuals from the two regression models being bivariate Gaussian.

The parameter values obtained were α = 0.156, β = −0.505 mmol/l in CARDS. In ASCOT, the parameters were α = 0.6807, β = 0.8850 mmol/l in the randomized dataset, and α = 0.4805, β = 0.5813 mmol/l in the observational dataset. This transformation has the same motivation as the inverse normal transform used in some GWAS applications (9, 10), but the use of a parametric transform preserves the relationship between pre and on-treatment measures, thereby allowing the difference between the two, adjusted for pretreatment value, to be used as a response variable as was done in ASCOT or as simply the on-treatment adjusted for pretreatment value as in CARDS (these are equivalent). The effect sizes in discovery cohorts (CARDS and ASCOT) and the replication cohort (PROSPER) were scaled so that the residuals had unit variance, thereby allowing studies using different transforms to be combined.

DNA extraction and genotyping

CARDS. DNA was extracted from whole-blood EDTA samples. DNA was isolated from 10 ml of frozen blood using the Gentra Puregene DNA Isolation Kit from Qiagen (Cat. no. 158389). Briefly, RBC was lysed with an anionic detergent in the presence of a DNA stabilizer that limits the activity of intracellular DNases. White blood cells were collected by centrifugation at 2,000 *g* for 2 min. RNA was removed by treatment with RNase A. Protein was removed by salt precipitation (centrifugation at 2000 *g* for 5 min). Genomic DNA was recovered by precipitation with isopropanol and centrifugation at 2,000 *g* for 5 min, the DNA pellet was washed in 70% ethanol, air dried, and dissolved in hydration solution (1 mM EDTA, 10 mM Tris·Cl, pH 7.5). Purified DNA was stored at −20°C. DNA aliquots were genotyped at Perlegen Sciences using a proprietary SNP set comprising 599,164 SNPs. Of these, 243 SNPs that had discrepant map positions between HapMap and Perlegen were dropped. We set a minimum SNP call rate threshold of 80% for including SNPs in the analysis, and we required that the *P*-value for a test of deviation from Hardy-Weinberg equilibrium (HWE) was not $< 10^{-5}$. This gave 517,746 SNPs for analysis. The average call rate was 98%, with 86.25% SNPs with a call rate of greater than 90%. SNP annotation was based on build 36 of the Human Genome Sequence. All SNPs were used in the analysis regardless of allele frequency, but the allele frequency was considered when evaluating putative associations. Allele frequency was below 1% at 6% of SNPs. We selected samples from those people who had been allocated atorvastatin 10 mg daily, had given consent for genotyping, and had a sample SNP call rate > 80%. After applying the exclusions of HWE, we estimated relatedness with PLINK, and those individuals with Pi_HAT > 0.25(excluding first- and second-degree relatives) were removed (n = 0). Only LDL-c values from time points at which the person was compliant with atorvastatin (based on pill count > 80%) were used.

ASCOT genotyping. Genotyping was carried out on HumanCNV370 (Illumina) array on 3,868 individuals at Centre National de Génotypage (CNG) in two batches. Samples were excluded if they had ⩾ 5% missing data (two samples). SNPs were excluded based on the following criteria: *i*) they had been mapped to different chromosomes or positions in the different releases (two SNPs), or *ii*) they were polymorphic A/T or C/G in either release or in the combined dataset, or *iii*) they had call rate ⩽ 97% in either release or in the combined dataset (47,744 SNPs), or *iv*) they had HWE *P*-value $\leqslant 10^{-7}$ in either release or in the combined dataset (8,502 SNPs). After applying the above exclusions, samples were excluded if they had estimated relatedness > 0.1875 (halfway cut point between second- and third-degree relatives), estimated using a using a subset of 101,954 SNPs obtained by linkage disequalibrium (LD)-based pruning (87 duplicates, 15 first-degree relatives and 4 presumed second-degree relatives removed. Then SNPs were excluded if they showed significant differences in allele frequency between the different batches at $P < 10^{-7}$ (20 SNPs), if they were monomorphic in the combined dataset (3,838 SNPs), if they were not in HapMap r22 (12,817 SNPs) or had different alleles to HapMap r22 (6 SNPs), or if they showed significant differences ($P < 10^{-7}$ using Fisher's exact test) in allele frequency between the combined dataset and HapMap r22 (308 SNPs). After applying all the above exclusions, ancestry outliers were excluded (n = 143) by using ancestry principal component analysis (11) on a subset of 100,905 SNPs selected by LD-based pruning, and ancestry principal components (PCs) were calculated for the remaining 3,804 individuals.

PROSPER genotyping. A whole genome-wide screening was performed in the sequential PHASE project. DNA was available for genotyping 5,763 subjects. Genotyping was performed with the Illumina 660K beadchip. After QC (call rate < 95%), 5,244 subjects and 557,192 SNPs were left for analysis (12).

Statistical analysis

Imputation of genotypes. The CARDS genotype data were combined with phased haplotypes from HapMap phase II CEU r22 to compute posterior probability distribution of genotype at all HapMap loci using the IMPUTE program (13). For ASCOT and PROSPER, genotypes at unmeasured SNPs were imputed using MACH (14) and phased haplotypes from HapMap CEU r22. For ASCOT, a randomly chosen subset of 400 individuals was used to estimate transition and emission probabilities (i.e., to estimate recombination rates between SNPs and per-SNP genotyping error rates) using MACH options "-greedy -r 100" for each (entire) chromosome in turn. Using these estimated rates (the .rec and .erate files), genotypes were imputed for the whole sample of 3,804 individuals using MACH options "-greedy-mle-mldetails" for each (entire) chromosome in turn.

CARDS data analysis. The EIGENSTRAT program (15) was used to adjust for population structure. Using PLINK (16), we generated a pruned subset of 152,587 SNPs that are in approximate linkage equilibrium with each other in the CARDS dataset. Principal components analysis was undertaken using this subset of SNPs. Thirty-seven individuals identified as outliers in the initial principal components analysis were excluded from the subsequent computation of principal components, leaving 1174 persons evaluable for statin response. The first three principal

components were retained and included as covariates in all tests of association.

On-treatment values for LDL-c for each individual at 1, 2, 3, 6, and 12 months post-randomization were available. We initially used the first available post-randomization LDL-c and established that the previously reported APOE genotype at rs445925 was the strongest association in a genome-wide analysis of response at $P = 1.1 \times 10^{-13}$. To maximize the power to detect any further new associations, we trained the weighting of post-randomization LDL-c time points to maximize the strength of the association of LDL response with *APOE* genotype at rs445925. Based on this, the nonmissing values for each individual were combined in a weighted average, with the one-month value allocated a weight of 0.6 and the four subsequent values, weights of 0.1 each (*P*-value for rs445925 with these weights = 2.2×10^{-16}). SNPTEST (13) was used to test for association of LDL response with genotype in a linear regression with the weighted average post-randomization LDL value as dependent variable and with covariates, including transformed pretreatment LDL-c, age, sex, and scores on the first three principal components of population stratification. The missing-data likelihood option was used to allow for uncertainty of genotypes at each imputed locus. In practice, the use of several weighted post-randomization LDL-c values rather than a single first value made very little difference to the results (see supplementary table II).

We used the conditional analysis test in PLINK (16) to test for independence of SNP associations over short regions within the same gene; a null model based on equating the effects of haplotypes that differed only at the SNP under test was compared with a more general model in which the effects of these haplotypes were unconstrained. The null hypothesis is that the SNP under test accounts for all associations of haplotypes with response. Other analyses included those carried out to explore initial associations, including a test of whether LPA genotype modifies the effect of atorvastatin on CVD. This was carried out by estimating the hazard ratio associated with allocation to atorvastatin in a Cox regression model of time to first CVD event and using a likelihood ratio test comparing a model with this main treatment effect and one including a term for interaction of genotype × treatment effect.

ASCOT data analysis. We regressed the response variable (transformed on-treatment minus transformed pretreatment LDL-c) onto imputed expected genotype dosage as implemented in ProbABEL (14, 17). This is asymptotically equivalent to score test for taking into account uncertainty in imputed genotypes (as in SNPTEST) but with improved finite sample size operating characteristics (18). Age, sex, age*sex, and transformed pretreatment LDL were used as covariates, plus 10 ancestry principal components.

PROSPER data analysis. The response variable was regressed (natural log of transformed on-treatment minus natural log of pretreatment LDL-c) onto imputed expected genotype dosage as implemented in SNPTEST. Age, sex, transformed pretreatment LDL, and top three principal components were used as covariates.

Meta-analysis. The score and observed information for the effect parameter were summed over studies to obtain a summary score test. This is algebraically equivalent (based on the quadratic approximation of the log-likelihood) to obtaining a weighted average of the maximum likelihood estimates with weights inversely proportional to the squared standard errors, with the useful feature that the ratio of observed to complete information (calculated by summing numerators and denominators over the three studies) is obtained as a summary measure of the efficiency of genotype imputation. For concise presentation, we focus here on showing the results of the meta-analysis rather than each study separately and provide study-specific estimates of effect only at the most extreme significance levels. In the data presentation, those loci at which the overall proportion of information extracted was less than 30% across the studies have been excluded. We have used the *P*-value threshold of $<5 \times 10^{-8}$ as the threshold for declaring a genome-wide significant association.

Distinguishing indirect and direct effects of genotype on on-treatment LDL. Effects of genetic variation on treatment response as measured by on-treatment LDL-c could be mediated through effects on the pretreatment LDL-c. To evaluate whether genetic on-treatment LDL-c likely reflects residual effect on pretreatment LDL-c, it is necessary to adjust for the pretreatment LDL-c levels and to correct the maximum likelihood estimate of the adjusted effect of genotype on on-treatment value for the noise in pretreatment values (the noise is both random measurement error and intra-individual variation in usual LDL-c). From the rules of path analysis, we calculated the direct effect γ of genotype on an on-treatment trait value as $\beta - \alpha\delta\ (1 - \rho) / \rho$, where β is the coefficient of regression for on-treatment trait value on genotype adjusted for measured pretreatment value, ρ is the intraclass correlation between replicate measurements of pretreatment values, and δ is the coefficient of regression for on-treatment value on observed pretreatment value. For these calculations, we used $\rho = 0.8$ as a plausible value for the intraclass correlation based on the within-person correlation in LDL-c values taken over two pretreatment visits in CARDS.

RESULTS

Table 1 compares baseline characteristics of participants in the three studies. **Fig. 1** shows a quantile-quantile plot of the $-\log_{10}$ *P*-values for association of each SNP with LDL-c response to treatment, obtained by meta-analyzing effect size estimates across the CARDS and ASCOT datasets. This plot shows that the cumulative distribution of test statistics approximates the null distribution over most of its range but that there is a tail of extreme results. **Fig. 2** shows a Manhattan plot of the $-\log_{10}$ *P*-values by map position. **Table 2** shows all loci at which the summary test for association yielded a nominal *P*-value $< 10^{-6}$. The estimates of effect (β) are for the transformed response variable (see Materials and Methods). In CARDS, the response variable was transformed on-treatment LDL-c with transformed pretreatment LDL-c entered as a covariate in the model. This is mathematically equivalent to modeling change in LDL-c with pretreatment LDL-c as a covariate (i.e., the difference in transformed on-treatment and adjustment for pretreatment LDL) as was done in ASCOT. A negative β for an allele means that the modeled allele is associated with a bigger reduction in posttreatment LDL-c and a better response to statins.

The strongest associations were with rs10455872 in the *LPA* gene on chromosome 6, and with SNPs in the *BCAM/PVRL2/APOE/APOC1* gene region on chromosome 19, where genome-wide significant associations were found. The SNPs in the *LPA* and *APOE* region explained 4% of the variance in LDL-c response in CARDS. The next

TABLE 1. Characteristics of patients and studies included in the meta-analysis

	CARDS	ASCOT-R	ASCOT-Obs
	n = 1194	*n = 895*	*n = 691*
Age (mean years ± SD)	61.6 ± 8.2	64.1 ± 8.0	64.2 ± 8.6
Ethnicity	Caucasian (UK and Ireland)	Caucasian (UK and Ireland)	Caucasian (UK and Ireland)
Women (%)	47	11.0	13.1
Diabetes (%)	100	21	21
Follow-up years (median IQR)	3.9 years (3.0–4.7)	First year was used	First year was used
Hypertension (%)	87	100	100
LDL-c level at baseline (mean mmol/l ± SD)	3.04 ± 0.71	3.47 ± 0.70	3.75 ± 0.85[a]
Lipid entry criterion	Fasting LDL-c ⩽ 4.14 mmol/l	Non-fasting TC ⩽ 6.5 mmol/l	None
Fasting status for lipids[b]	Overnight fast	Fasting	Fasting
Statin dose	Atorvastatin 10 mg daily	Atorvastatin 10 mg daily	Atorvastatin 10 mg daily
Platform	Perlegen 6	Illumina HumanCNV370	Illumina HumanCNV370
pHWE[c] exclusion	10^{-5}	10^{-7}	10^{-7}
Imputation software	IMPUTE 2	MACH	MACH
NCBI build for imputation	HapMap CEU r22	HapMap CEU r22	HapMap CEU r22

ASCOT-Obs, observational arm of ASCOT; ASCOT-R, randomized arm of ASCOT.
[a]In N = 656 with nonmissing LDL-c at baseline; the missingness is nonrandom because these are individuals with baseline triglycerides too high for Friedewald formula.
[b]Fasting status for LDL-c at baseline (see previous row) and for response to statin measure.
[c]*P*-value threshold for exclusion of SNPs not in HWE.

most significant *P*-value was that for the *ALG10* region on chromosome 12, but this did not reach genome-wide significance. There was no evidence of gene-gender interaction for all the top SNPs reported in the study. The effect sizes for all top SNPs were similar in CARDS, where all the participants had type 2 diabetes, and in ASCOT, where 21% of the participants had type 2 diabetes (Table 2), suggesting that diabetes per se was not a strong determinant of the genetic effect of the top SNPs.

LPA

In the *LPA* gene SNP, rs10455872 showed a genome-wide significant association with LDL-c response (**Fig. 3**). The effect at rs10455872 was modest; the β shown in Table 2 is not easily directly interpretable given the transformation used, but in CARDS, for example, the percentage change in LDL-c with statin therapy at one month was approximately −43% in those with at least one "G" allele at rs10455872 (MAF = 8%) compared with −46.5% in homozygotes for the "A" allele. There was no significant effect of this SNP on change in LDL-c post-randomization in those in the placebo group ($P = 0.28$).

To investigate the association with *LPA* genotype further, we first confirmed that *LPA* genotypes predicted serum lipoprotein(a) [Lp(a)] levels, which had been measured in CARDS but not in ASCOT. **Fig. 4** shows the results of the GWAS for serum Lp(a) levels; all significantly associated loci were in the *LPA* region, consistent with other reports (19). In a linear regression model that included age, sex, and population structure covariates, 12 SNPs in the *LPA* region had independent effects on serum Lp(a) (rs10455872, rs5014650, rs783147, rs6919346, rs3103349,

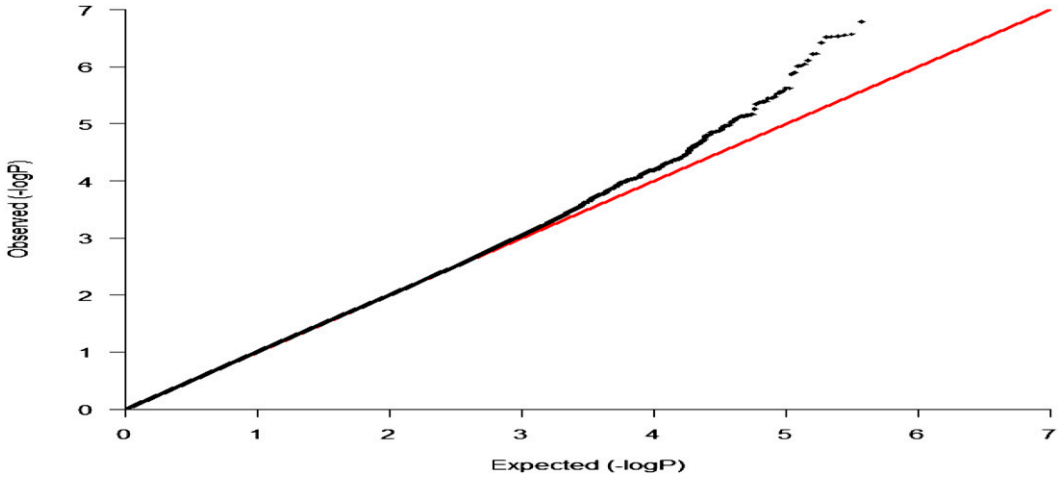


Fig. 1. Quantile-quantile plot of meta-analysis *P*-values for statin response. A plot of the quantiles of observed and expected distribution of *P*-values against each other.

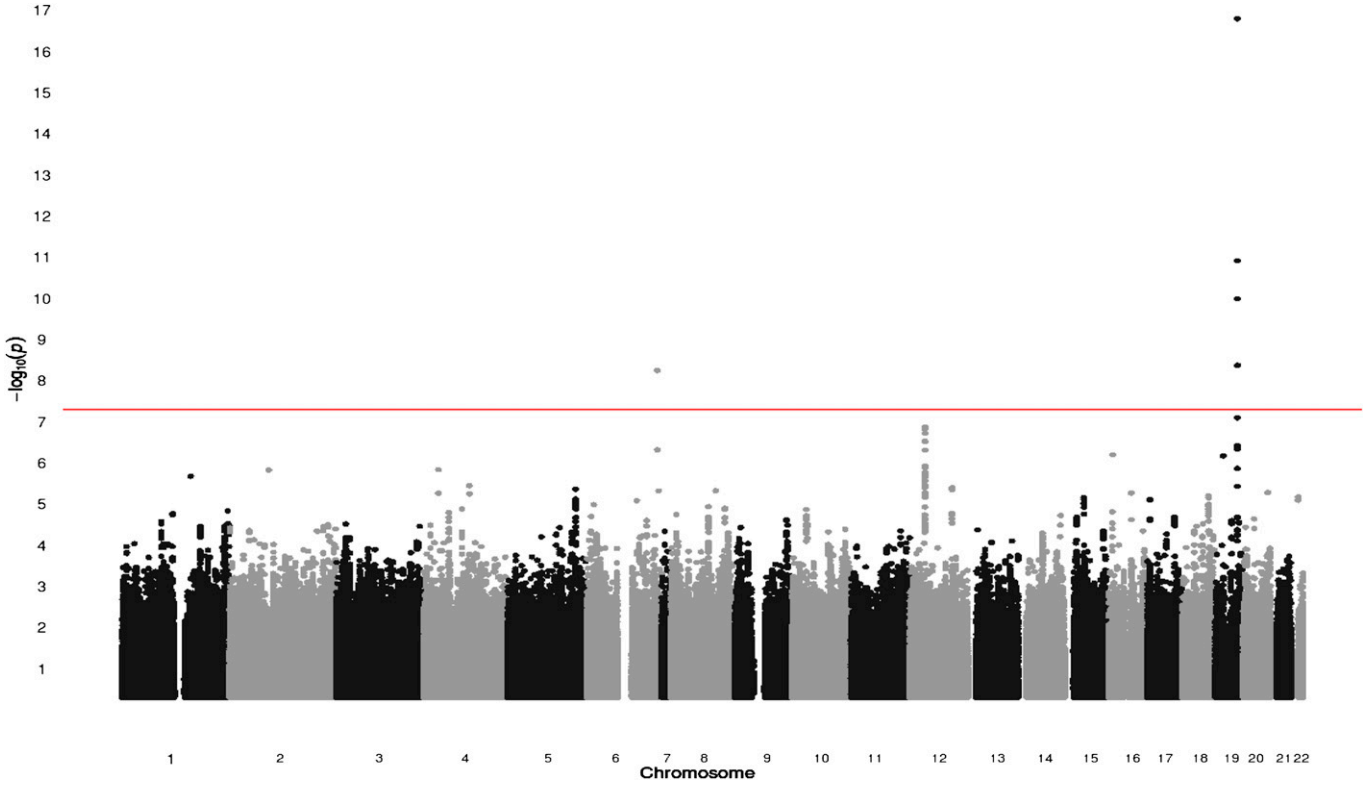


Fig. 2. Manhattan plot of *P*-values from meta-analysis of all SNPs that passed stringent quality control. The Manhattan plots [also known as $-\log_{10}(P)$ association plots[show the chromosomal position of SNPs exceeding the genome-wide significance threshold ($P < 5 \times 10^{-8}$) as indicated by the solid red line.

rs2063347, rs6415084, rs10455782, rs394487, rs6926458, rs316174, rs3127569). Together, these SNPs explained 40% variation in the serum Lp(a) levels in CARDS; however, most of this was attributable to the rs10455872 SNP (30%) with median levels being 7.6 mg/dl [interquartile range (IQR) 4.1–14 mg/dl], 50.5 mg/dl (IQR 37–68 mg/dl) and 55.2 mg/dl (IQR 51–113 mg/dl) in those with AA, AG, and GG genotype, respectively. We then adjusted the

TABLE 2. Combined analysis (CARDS, ASCOT randomized, and ASCOT observational)

CHR	POS (cM)	SNP	Modeled Allele	Minor Allele (Frequency)	CARDS β	CARDS SE	Ascot-R β	Ascot-R SE	Ascot-Obs β	Ascot-Obs SE	Meta-analysis Rsq[a]	Meta-analysis β[b]	Meta-analysis SE	*P*	Gene[c]
6	195.419	rs10455872	A	G (0.07)	−0.35	0.08	−0.36	0.11	−0.1	0.18	0.54	−0.35	0.06	6.13E-09	LPA
12	55.598	rs1627770	G	T (0.2)	0.18	0.05	0.13	0.05	0.17	0.06	1	0.17	0.03	1.81E-07	LOC390301-ALG10
12	55.598	rs863626	C	T (0.2)	0.18	0.05	0.13	0.05	0.17	0.06	1	0.18	0.03	1.39E-07	LOC390301-ALG10
12	55.598	rs11053045	A	T (0.2)	0.18	0.05	0.13	0.05	0.17	0.06	1	0.18	0.03	1.34E-07	LOC390301-ALG10
12	55.598	rs1619785	A	A (0.2)	−0.18	0.05	−0.13	0.05	−0.17	0.06	1	−0.18	0.03	1.28E-07	LOC390301-ALG10
12	55.599	rs10844779	A	A (0.2)	−0.18	0.05	−0.14	0.05	−0.17	0.06	1	−0.18	0.03	1.44E-07	ALG10-LOC260338
12	55.599	rs11053068	C	C (0.2)	−0.18	0.05	−0.14	0.05	−0.17	0.06	1	−0.18	0.03	1.44E-07	ALG10-LOC260338
12	55.599	rs5004272	A	G (0.21)	0.18	0.05	0.12	0.05	0.17	0.06	1	0.17	0.03	2.81E-07	ALG10-LOC260338
12	55.599	rs10844823	C	C (0.21)	−0.18	0.05	−0.12	0.05	−0.17	0.06	0.99	−0.17	0.03	2.86E-07	ALG10-LOC260338
16	27.656	rs721843	C	G (0.46)	0.13	0.04	0.12	0.05	0.14	0.05	0.97	0.14	0.03	6.05E-07	LOC653737-GRIN2A
19	80.713	rs4803760	C	T (0.2)	0.24	0.06	0.08	0.05	0.19	0.07	0.97	0.18	0.04	4.23E-07	BCAM-PVRL2
19	80.766	rs1985096	A	A (0.16)	−0.33	0.07	−0.16	0.06	−0.28	0.08	0.8	−0.27	0.04	9.49E-11	BCAM-PVRL2
19	80.877	rs395908	A	A (0.16)	−0.21	0.06	−0.1	0.06	−0.23	0.07	0.92	−0.19	0.04	3.46E-07	PVRL2-BCAM-TOMM40
19	80.954	rs6857	C	T (0.14)	−0.32	0.07	−0.06	0.07	−0.23	0.08	0.93	−0.23	0.04	7.43E-08	PVRL2-BCAM-TOMM40
19	81.023	rs405509	G	T (0.48)	−0.17	0.05	−0.1	0.04	−0.21	0.05	0.99	−0.17	0.03	3.46E-09	APOE-TOMM40-APOE
19	81.051	rs445925	A	A (0.11)	−0.44	0.08	−0.36	0.07	−0.34	0.09	0.77	−0.42	0.05	1.59E-17	LOC100129500-APOE APOC1
19	81.081	rs4420638	A	G (0.16)	−0.44	0.08	−0.15	0.07	−0.32	0.09	0.56	−0.33	0.05	1.12E-11	APOC1 APOC1 APOC4

SNPs associated with LDL-c response to statins with meta-analysis values of $P < 10-^{6}$ and Rsq > 0.30.
Ascot-Obs, ASCOT observational; Ascot-R, ASCOT randomized; CHR, chromosome; POS, position.
[a]Estimate of squared correlation between imputed and true genotypes.
[b]A positive β value means that the modeled allele is associated with a bigger posttreatment LDL-c and, therefore, a lower response to statins. A negative β value means that the modeled allele is associated with lower posttreatment LDL-c and, therefore, a better response to statins.
[c]For SNPs that lie in the intergenic regions, the location of the nearby genes is shown.

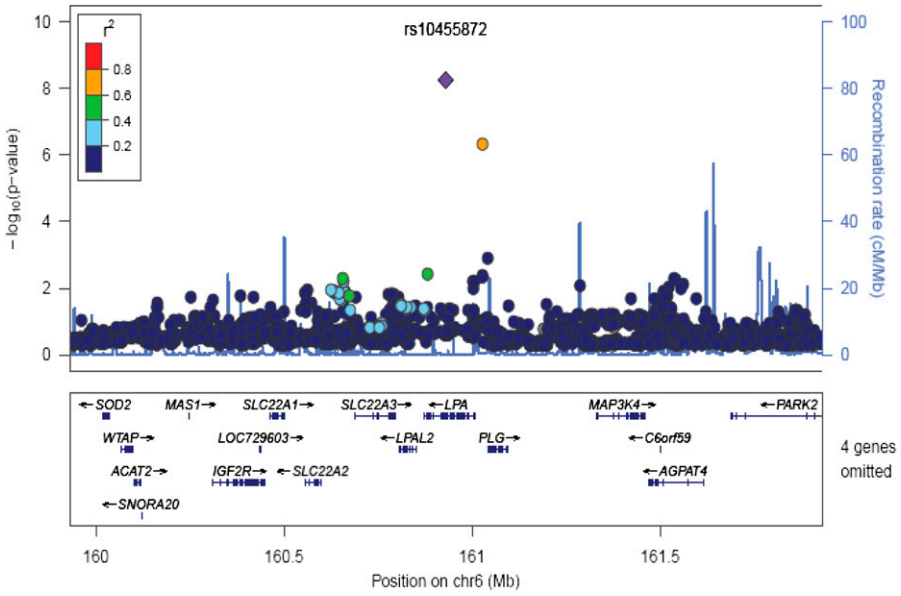


Fig. 3. Regional association plot of LPA locus with statin response. Correlations between the target SNP (the SNP with the lowest *P* value, depicted in purple) and nearby SNPs within a 500 kb region. The r^2 values were based on the HapMap CEU population.

association of genotype at SNPs in the *LPA* gene with LDL-c response to statin for measured serum Lp(a) levels in the CARDS data to test whether the genetic effects seen are likely to be mediated through the effect of *LPA* on serum Lp(a) levels (**Table 3**). The estimate of the standardized regression coefficient at the associated SNP (rs10455872) in *LPA* in CARDS was reduced from −0.35 (±0.08) to −0.09 (±0.08), consistent with the effect of genotype on apparent response to statin being mediated through Lp(a) levels. We noted that Lp(a) levels had an independent association with apparent LDL-c response to statin beyond genotype in these analyses ($P = 0.001$). Further analysis in CARDS also confirmed that there was no effect of statin on Lp(a) levels; β was −0.23 mg/dl, (95% CI: −2.25 to 1.80) for difference in Lp(a) levels with atorvastatin versus placebo at one year post-randomization, adjusted for baseline Lp(a), age, and sex.

To assess whether serum Lp(a) levels might alter efficacy of statin therapy on CVD itself, we examined whether there was any evidence of interaction (deviation from a multiplicative model of joint effects on a hazard scale) between high serum Lp(a) levels and atorvastatin on CVD end points in CARDS. The hazard ratio for CVD events associated with statin use was 0.60 (95% CI: 0.32–1.13) among those in the top quartile for serum Lp(a) (>22 mg/dl) compared with 0.66 (95% CI: 0.46–0.93) among those

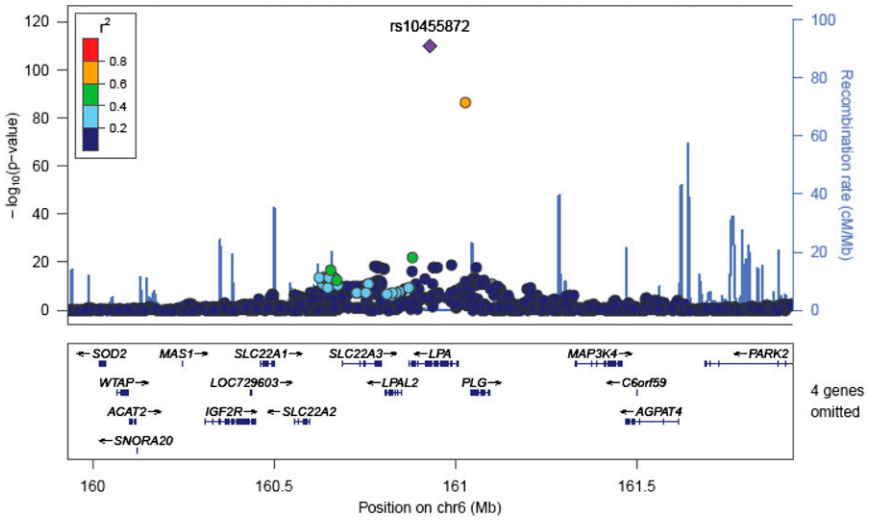


Fig. 4. Regional association plot of LPA locus with Lp(a) levels in the CARDS dataset. Correlations between the target SNP (the SNP with the lowest *P* value, depicted in purple) and nearby SNPs within a 500 kb region. The r^2 values were based on the HapMap CEU population.

TABLE 3. Effect of adjustment for serum Lp(a) levels in CARDS

CHR	POS (cM)	SNP	Modeled Allele	Minor Allele (Frequency)	Before Lp(a) Adjustment β	SE	P	After Lp(a) Adjustment β	SE	P	Gene[a]
6	195.419	rs10455872	A	G (0.07)	−0.35	0.08	1.12E-05	−0.09	0.08	2.96E-01	LPA
12	55.598	rs1627770	G	T (0.2)	0.18	0.05	5.08E-04	0.19	0.05	3.21E-04	LOC390301-ALG10
12	55.598	rs863626	C	T (0.2)	0.18	0.05	3.83E-04	0.19	0.05	2.48E-04	LOC390301-ALG10
12	55.598	rs11053045	A	T (0.2)	0.18	0.05	3.73E-04	0.2	0.05	2.40E-04	LOC390301-ALG10
12	55.598	rs1619785	A	A (0.2)	−0.18	0.05	3.60E-04	−0.2	0.05	2.28E-04	LOC390301-ALG10
12	55.599	rs10844779	A	A (0.2)	−0.18	0.05	4.22E-04	−0.19	0.05	2.79E-04	ALG10-LOC260338
12	55.599	rs11053068	C	C (0.2)	−0.18	0.05	4.27E-04	−0.19	0.05	2.82E-04	ALG10-LOC260338
12	55.599	rs5004272	A	G (0.21)	0.18	0.05	5.82E-04	0.19	0.05	4.14E-04	ALG10-LOC260338
12	55.599	rs10844823	C	C (0.21)	−0.18	0.05	5.88E-04	−0.19	0.05	4.19E-04	ALG10-LOC260338
16	27.656	rs721843	C	G (0.46)	0.13	0.05	4.67E-01	0.13	0.04	3.69E-01	LOC653737-GRIN2A
19	80.713	rs4803760	C	T (0.2)	0.24	0.06	3.43E-05	0.23	0.06	7.73E-05	BCAM-PVRL2
19	80.766	rs1985096	A	A (0.16)	−0.33	0.07	8.31E-07	−0.33	0.07	1.39E-06	BCAM-PVRL2
19	80.877	rs395908	A	A (0.16)	−0.21	0.06	2.10E-04	−0.21	0.06	3.65E-04	PVRL2-BCAM-TOMM40
19	80.954	rs6857	C	T (0.14)	−0.32	0.07	1.85E-06	−0.3	0.07	1.75E-05	PVRL2-BCAM-TOMM40
19	81.023	rs405509	G	T (0.48)	−0.17	0.05	3.36E-04	−0.14	0.05	4.31E-03	APOE-TOMM40-APOE
19	81.051	rs445925	A	A (0.11)	−0.44	0.08	1.13E-08	−0.42	0.08	1.39E-07	LOC100129500-APOE APOC1
19	81.081	rs4420638	A	G (0.16)	−0.44	0.08	1.65E-08	−0.43	0.08	1.39E-07	APOC1 APOC1 APOC4

Effect of adjustment for serum Lp(a) levels in CARDS for SNPs associated with an LDL-c response to statins with a meta-analysis of $P<10^{-6}$. CHR, chromosome; POS, position.

[a]For SNPS that lie in the intergenic regions, the location of the nearby genes is shown.

with serum Lp(a) below this level (likelihood ratio test for interaction $P = 0.8$).

Nor was there any evidence of interaction between rs10455872 genotype at *LPA* and atorvastatin for effects on CVD end points ($P = 0.27$ for the interaction of genotype at rs10455872 locus). Here the HR associated with atorvastatin in those homozygous for the A allele was 0.58 (95% CI: 0.41–0.83) and the HR in those with at least one G allele was 1.03 (95% CI: 0.38–2.78). However, the power to detect such an interaction was limited as there were only 16 events among the 294 trial participants with at least one copy of the G allele.

Replication of the LPA SNP

We tested the effect of *LPA* SNP (rs10455872) in 2,550 participants in the PROSPER trial randomized to 40 mg/day of pravastatin. In this study, "A" allele of rs10455872 was also associated with lower response to statins with a scaled β of –0.18 ± 0.04, $P = 0.009$. The combined *P*-value for the three studies was 1.2E-09 (β = −0.28 ± 0.04).

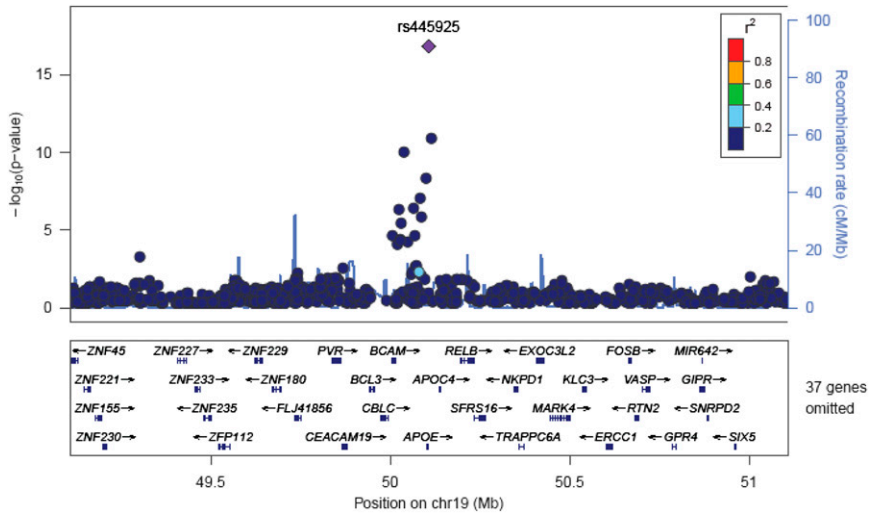


Fig. 5. Regional association plot of APOE locus with statin response. Correlations between the target SNP (the SNP with the lowest *P* value, depicted in purple) and nearby SNPs within a 500 kb region. The r^2 values were based on the HapMap CEU population.

APOE

Several SNPs in the *BCAM/PVRL2/APOE/APOC1/APOE* gene region reached genome-wide significance for statin response (**Fig. 5**). The effect on LDL-c response to statin therapy associated with these SNPs at in this region was modest; in CARDS for example the % change in LDL-c with statin therapy at one month was approximately -51% in those with at least one "A" allele at rs445925 compared with -45% in common GG homozygotes and was approximately -37% in those with at least one "G" allele at rs4420638 compared with -47% in common "AA" homozygotes. These effects were independent of the effect of genotype at rs10455872 in *LPA* and of Lp(a) levels. In the CARDS dataset we confirmed that there was no significant effect of these SNPs on change in LDL-c post randomization in those in the placebo group ($P = 0.47$).

We examined whether the effects in this region could be accounted for by the known $\varepsilon2/\varepsilon3/\varepsilon4$ protein polymorphism of apolipoprotein E, which corresponds to *APOE* SNP haplotypes T-T, T-C, and C-C, respectively, at rs429358 and rs7412. The presence of the "T" allele at rs7412 contrasts the $\varepsilon2$ protein variant with other protein variants, whereas presence of the "C" allele at rs429358 contrasts the $\varepsilon4$ protein variant with other protein variants. These two SNPS were not directly typed and could not be imputed as they are not in the HapMap II.

The "A" allele at rs445925, which we found to be associated with a higher statin response ($\beta = -0.44$) is in LD with the "T" allele at rs7412 with a reported r^2 of 0.76; thus, it is a proxy for the $\varepsilon2$ protein variant (20). The "G" allele at rs4420638, which was associated with lower response to statin, is in LD with the "C" allele at rs429358 with reported r^2 of 0.62 but with a low r^2 of 0.01 for rs7412 (21); thus, it is a proxy for the $\varepsilon4$ protein variant.

These two proxy SNPs are in the HapMap and could be imputed in this analysis with percentage information content (i.e., imputation quality) of 77% and 56%, respectively. Thus we tested for residual effects of SNP haplotypes conditioning either on rs445925 (as a proxy for rs7412) or on rs4420638 (as a proxy for rs429358). When conditioned on rs4420638, the $\varepsilon4$ proxy, the additional percentage variance explained by residual haplotype effects is 0.7% (F statistic with 8 and 854, df = 3.21, $P = 0.001$). When conditioned on rs445925, the proxy for $\varepsilon2$, the additional percentage variance explained by residual haplotype effects is only 0.2% (F statistic with 8 and 854, df = 1.76, $P = 0.08$), suggesting that $\varepsilon2$ accounts for most of the variance in response at this locus.

ALG10

Beyond these associations of LDL-c response with *APOE* and *LPA*, no other genome-wide significant associations were found. The next most significant SNPS were those in the ALG10 gene region (**Fig. 6**) on chromosome 12 where several SNPs had $P < 10^{-6}$. ALG10 codes for asparagine-linked glycosylation protein 10 homolog A. Of these SNPs, most map to intergenic regions either side of the ALG10 gene itself with one imputed SNP within ALG10 having a *P*-value for association with statin response of 6.79×10^{-6}.

Effect of pretreatment LDL-C

To demonstrate that these findings are unlikely to be confounded by baseline LDL-c, **Table 4** shows unadjusted, adjusted, and corrected estimates of the direct effect of genotype on posttreatment LDL at the strongest SNPs for the *APOE* region *LPA* and *ALG10* in the CARDS dataset. At the *APOE* $\varepsilon2$ proxy SNP (rs445925), without adjusting for baseline LDL-c, the apparent LDL-c response to statins would be more than double that observed in our baseline-adjusted model ($\beta = -1.01$ vs. -0.44 per copy of "A" allele), emphasizing the effect of adjusting for baseline LDL-c. However, adjusting our effect size estimate further by modeling measurement noise at baseline reduced the apparent effect just slightly to $\beta = -0.30$, suggesting there

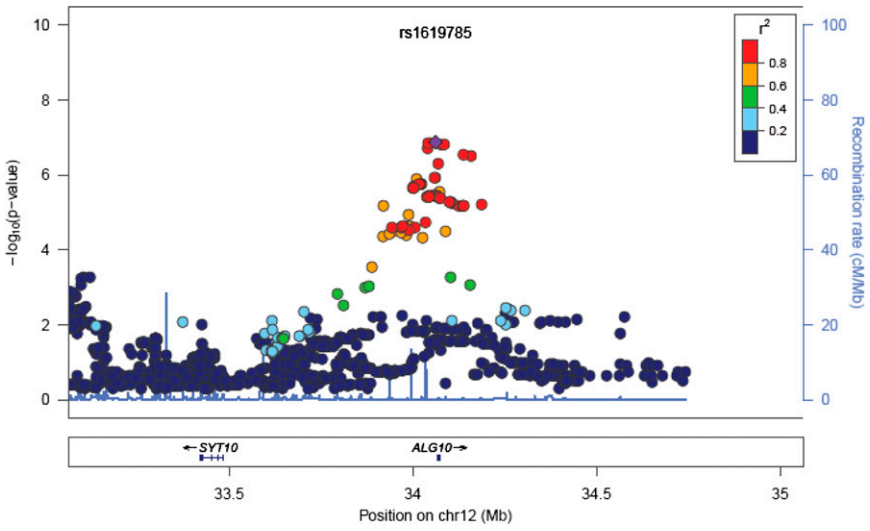


Fig. 6. Regional association plot of ALG10 locus with statin response before Lp(a) adjustments. Correlations between the target SNP (the SNP with the lowest *P* value, depicted in purple) and nearby SNPs within a 500 kb region. The r^2 values were based on the HapMap CEU population.

TABLE 4. Effect of genotype on posttreatment LDL-c (CARDS only)

SNP	β unadjusted for baseline LDL	β adjusted for observed baseline LDL but uncorrected for measurement noise	β adjusted for baseline LDL and corrected for measurement noise
rs445925	−1.01	−0.44	−0.38
rs4420638	−0.54	−0.44	−0.42
rs10455872 (*LPA*)	−0.49	−0.35	−0.32
rs10844779 (*ALG10*)	−0.18	−0.18	−0.18

With and without correction for measurement noise in baseline LDL.

is little residual effect of baseline due to measurement noise. At the *APOE* ε4 proxy SNP and at the *LPA* SNP, the estimated effect of baseline LDL-c adjustment is much less, and thus, the adjustment for measurement noise alters the association only slightly.

Other genes of interest

Previously reported variants associated with statin response in the *PCSK9* (rs11591147), *HMGCR* (rs1047443, rs17671591, rs6453131), *KIF-6* (rs20455), *ABCB1* (1236/2677/3435 TTT haplotype), *CLMN* (rs80141914, associated with TC response to statin), and *GCKR* (rs1260326 associated with triglyceride level response to statin) genes were not significantly associated at with LDL-c response to statin in this study at an accepted genome-wide association threshold ($P \leqslant 10^{-8}$) or even at thresholds typically expected to declare replication (say, $P \leqslant 10^{-2}$). However, *PCSK9* (rs11591147) and *GCKR* (rs1260326) were significant at a threshold of 0.05 (see supplementary Table I). We have refrained from comparing the directionality and magnitude of these effects in the present study because of the different phenotype characterization and transformations across the studies and, in some studies, lack of information about the modeled alleles.

DISCUSSION

In this genome-wide association study of LDL-c response to atorvastatin therapy, we report that those with genotypes in the *LPA* gene that lead to higher Lp(a) levels have an apparently lower LDL-c response to statin, and we replicate the previously reported association of a higher response to statin in those with the A allele at the *APOE* ε2 locus. The top three SNPs in the study, rs10455872 in *LPA* and the *APOE* ε2 and *APOE* ε4 variants, explained only 4% variance in the LDL-c response to statin treatment; however, it is possible that that larger studies might detect more SNPs with smaller effect sizes or that there are larger effects at rarer variants not captured by our imputed genotypes.

LPA

Lipoprotein(a) is a plasma lipoprotein consisting of a cholesterol-rich LDL particle with one molecule of apolipoprotein B100 and an additional protein, apolipoprotein(a), attached via a disulfide bond. Serum levels of Lp(a) have a highly skewed distribution; for example, in CARDS, the median serum Lp(a) was 8.9 mg/dl (IQR 4.5–21.3 mg/dl) and with values as high as 238 mg/dl. Approximately 30% of variance in Lp(a) levels has been reported as determined by the kringle IV type 2 (KIV-2) copy number variant in *LPA*, which is known to encode variability in the size of apo(a). Some variance in measured Lp(a) attributable to genes is also due to apo(a) size heterogeneity affecting the results of the immunochemical methods used to quantify Lp(a), as is the case with the assay we used (22). That is, genotype can induce some measurement error in Lp(a), although recent data from the Framingham study suggest the measurement error is likely to be of little practical importance (23). The Lp(a)-raising genotype associated with the kringle repeat and high Lp(a) levels themselves have also been reported to be associated with increased cardiovascular risk in several studies (24–28). As such, recent guidelines emphasize the importance of detecting high Lp(a) phenotype and possible intervention with niacin (26).

The rs10455872 SNP that we found associated with LDL-c response is in strong LD with the KIV-2 copy number variant in Lp(a) (29). Consistent with this, variation at rs10455872 accounted for 30% of variance in Lp(a) in the CARDS data. However, the explanation for the apparently lower LDL-c response in those with genotypes associated with high Lp(a) lies in understanding what LDL-c estimation actually captures. The standard Friedewald formula calculates LDL-c levels from TC, HDL-cholesterol, and plasma triglyceride and actually includes the cholesterol that resides in Lp(a). For most patients, this is of little importance as usually only about 5% of what is measured as LDL-cholesterol is estimated to reside in Lp(a). However, it is estimated that about 8% of apparent LDL-c resides in Lp(a) if Lp(a) levels are in the range 30–60 mg/dl and as much as 20% if Lp(a) is > 60 mg/dl (30). As we show definitively here in the CARDS trial, statin therapy did not lower Lp(a) levels. Thus, individuals who had an appreciable fraction of their total plasma cholesterol carried on Lp(a) particles had some cholesterol in statin-responsive LDL particles and some in statin-unresponsive Lp(a) particles. For such patients, true LDL-c response will be underestimated because apparent on-treatment LDL-c will comprise truly falling LDL-c but static Lp(a) levels. This phenomenon has previously been noted in the context of nephrotic syndrome (31) and has been emphasized by Scanu et al. (32). Our estimate that those with at least one copy of the Lp(a)-raising G allele at rs10455872 have about a 5 percentage points lower apparent statin response (45% in "GG" and "AG" genotype vs. 40% in "AA" genotype) and that this association disappears when adjusted for Lp(a) levels is consistent with these observations.

Although the effect of the G allele on statin response is modest, this allele only accounts for about 30% of variance in Lp(a) levels. The data highlight a more general clinical point that individuals with raised Lp(a) levels for any reason have a somewhat lower apparent response to statin therapy and, therefore, that an apparently lower LDL-c response to statin may be an indication for checking Lp(a)

levels. However, we also show here that similar relative protection from CVD with atorvastatin therapy was found in those with and without elevated Lp(a). It is increasingly accepted that elevated Lp(a) increases CVD risk (24, 25). Therefore, it is important that, although Lp(a) levels themselves are not changed and statin effects on LDL-c appear erroneously low, statin therapy be continued in individuals with high Lp(a).

We confirmed the association of LPA SNP rs10455872, that is, the association of "A" allele with a lower response to statins in an independent cohort of 2,550 subjects randomized to 40 mg of pravastatin. This is the first report of a successful replication of genetic response to statin treatment beyond the *APOE* region in a genome-wide association study.

APOE

We replicated the previous finding that genotype at the *APOE* ε2 locus is associated with variation in statin response. Having at least one "A" allele at rs445925, which is in strong LD with the locus determining the ε2 protein variant, was associated with both higher baseline LDL-c and greater response to statin, whereas the proxy for ε4 protein variant was associated with lower LDL-c response to statin. The conditional haplotype analysis suggests most of the variation is attributable to the number of ε2 copies rather than to the number of ε4 copies, but this is not definitive given that there is uncertainty in the haplotypes and that the HapMap SNPs are imperfect proxies. As noted previously, individuals in whom a higher proportion of cholesterol is synthesized rather than taken up via diet, such as ε2 carriers, are more susceptible to inhibition of cholesterol synthesis (2), and in addition, there may be more remnant and IDL-like particles contributing to apparent LDL-c in those patients with an *APO* ε2 allele. Statins are very good at removing these larger LDL particles through LDL receptor upregulation.

ALG10

We found some suggestion of an association between statin response and variants in the ALG10 gene region, but this did not reach the usual genome-wide significance threshold of $P < 10^{-8}$; therefore, it requires confirmation in other studies before considering it other than a spurious association. ALG10 codes for asparagine-linked glycosylation protein 10 homolog A, which adds the third glucose residue to the lipid-linked oligosaccharide precursor for N-linked glycosylation. Its relevance to statin response remains to be established, though of course, protein modification by N-linked glycosylation is relevant to diverse aspects of human biology, including functional modification of many enzymes (33).

Previous studies

Two GWAS of LDL-c response to statin therapy have previously been reported (2, 3). In the Treating to New Targets (TNT) trial dataset with 1,984 treated individuals typed with a genome-wide panel, there were no loci at which *P*-values for association were less than 10^{-6}, but in a superset of 5,745 individuals studied for candidate gene associations only, three SNPs in *APOE* and one SNP in *PCSK9* reached genome-wide significance (2). In a meta-analysis of three trials that included 3,932 treated subjects (3) a SNP in the *CLMN* gene was significant at $P < 10^{-7}$ for association with TC response, and there was a weak association with SNP in *APOE*. In the same study, polymorphism in the *GCKR* gene was shown to be associated with statin-induced change in triglycerides. Candidate gene analyses have shown that a common *LDLR 3-UTR* haplotype is associated with attenuated lipid-lowering response to simvastatin treatment (34). In the same study, *HMGCR* gene polymorphisms were also associated with reduced plasma LDL-c and with reduced LDL-c response to simvastatin. The association of *HMGCR* gene with statin response was also reported in a population-based cohort of patients with diabetes (35). These effects were more evident in African-Americans than in European-Americans. In a separate study, carried out in acute coronary syndrome patients, carriers of a polymorphism in *kinesin-like protein 6 (KIF-6)* have been reported to have greater benefit from pravastatin versus placebo with respect to CVD outcome but not with respect to lipid or C-reactive peptide response (36). Additionally, association of the *ABCB1* gene with statin response has been reported (4, 37). Apart from the *APOE* association, none of these other associations were replicated here.

Finally, we note that the effects identified in this study are of modest size: the importance of further studying the genetics of response to statin therapy may be not in predicting who will benefit from statins but in identifying other therapeutic targets.

The authors thank the other investigators, the staff, and the participants of the CARDS and ASCOT studies for their important contributions. Generated data for the study: Helen M. Colhoun, D. John Betteridge, John H. Fuller, Shona Livingstone, Valentine Charlton-Menys, Andrew Neil, Neil Poulter, Peter Sever, Denis Shields, Alice Stanton, Alu Chatterjee, Craig Hyde, Roberto A. Calle, David A. DeMicco, Graham A. Hitman, Mark Caulfield, Stella Trompet, Iris Postmus, Ian Ford, and J. Wouter Jukema. Analyzed data: Paul M. McKeigue, Toby Johnson, Harshal Deshmukh, Helen M. Colhoun, Graham Hitman, Mark Caulfield, Stella Trompet, and J. Wouter Jukema. Drafted initial manuscript: Helen M. Colhoun, Paul M. McKeigue, Harshal Deshmukh, Toby Johnson, Graham Hitman, and Mark Caulfield. Edited manuscript: All authors. See Refs. 1, 6, and 8 for a list of CARDS, ASCOT, and PROSPER investigators, respectively.

REFERENCES

1. Colhoun, H. M., D. J. Betteridge, P. N. Durrington, G. A. Hitman, H. A. Neil, S. J. Livingstone, M. J. Thomason, M. I. Mackness, V. Charlton-Menys, and J. H. Fuller. 2004. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. *Lancet.* **364:** 685–696.
2. Thompson, J. F., C. L. Hyde, L. S. Wood, S. A. Paciga, D. A. Hinds, D. R. Cox, G. K. Hovingh, and J. J. Kastelein. 2009. Comprehensive

whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort. *Circ. Cardiovasc. Genet.* **2:** 173–181.
3. Barber, M. J., L. M. Mangravite, C. L. Hyde, D. I. Chasman, J. D. Smith, C. A. McCarty, X. Li, R. A. Wilke, M. J. Rieder, P. T. Williams, et al. 2010. Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS ONE.* **5:** e9763.
4. Thompson, J. F., M. Man, K. J. Johnson, L. S. Wood, M. E. Lira, D. B. Lloyd, P. Banerjee, P. M. Milos, S. P. Myrand, J. Paulauskis, et al. 2005. An association study of 43 SNPs in 16 candidate genes with atorvastatin response. *Pharmacogenomics J.* **5:** 352–358.
5. Donnelly, L. A., C. N. Palmer, A. L. Whitley, C. C. Lang, A. S. Doney, A. D. Morris, and P. T. Donnan. 2008. Apolipoprotein E genotypes are associated with lipid-lowering responses to statin treatment in diabetes: a Go-DARTS study. *Pharmacogenet. Genomics.* **18:** 279–287.
6. Sever, P. S., B. Dahlof, N. R. Poulter, H. Wedel, G. Beevers, M. Caulfield, R. Collins, S. E. Kjeldsen, A. Kristinsson, and G. T. McInnes. 2003. Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial–Lipid Lowering Arm (ASCOT-LLA): a multicentre randomised controlled trial. *Lancet.* **361:** 1149–1158.
7. Friedewald, W. T., R. I. Levy, and D. S. Fredrickson. 1972. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **18:** 499–502.
8. Shepherd, J., G. J. Blauw, M. B. Murphy, S. M. Cobbe, E. L. Bollen, B. M. Buckley, I. Ford, J. W. Jukema, M. Hyland, A. Gaw, et al. 1999. The design of a prospective study of Pravastatin in the Elderly at Risk (PROSPER). PROSPER Study Group. PROspective Study of Pravastatin in the Elderly at Risk. *Am. J. Cardiol.* **84:** 1192–1197.
9. Lindgren, C. M., I. M. Heid, J. C. Randall, C. Lamina, V. Steinthorsdottir, L. Qi, E. K. Speliotes, G. Thorleifsson, C. J. Willer, B. M. Herrera, et al. 2009. Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.* **5:** e1000508.
10. Willer, C. J., E. K. Speliotes, R. J. Loos, S. Li, C. M. Lindgren, I. M. Heid, S. I. Berndt, A. L. Elliott, A. U. Jackson, C. Lamina, et al. 2009. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41:** 25–34.
11. Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. *PLoS Genet.* **2:** e190.
12. Trompet, S., A. J. de Craen, I. Postmus, I. Ford, N. Sattar, M. Caslake, D. J. Stott, B. M. Buckley, F. Sacks, J. J. Devlin, et al. 2011. Replication of LDL GWAs hits in PROSPER/PHASE as validation for future (pharmaco)genetic analyses. *BMC Med. Genet.* **12:** 131.
13. Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39:** 906–913.
14. Li, Y., C. Willer, S. Sanna, and G. Abecasis. 2009. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10:** 387–406.
15. Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38:** 904–909.
16. Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81:** 559–575.
17. Aulchenko, Y. S., M. V. Struchalin, and C. M. van Duijn. 2010. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics.* **11:** 134.
18. Kutalik, Z., T. Johnson, M. Bochud, V. Mooser, P. Vollenweider, G. Waeber, D. Waterworth, J. S. Beckmann, and S. Bergmann. 2011. Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics.* **12:** 1–17.
19. Zabaneh, D., M. Kumari, M. Sandhu, N. Wareham, N. Wainwright, T. Papamarkou, J. Hopewell, R. Clarke, K. Li, J. Palmen, et al. 2011. Meta analysis of candidate gene variants outside the LPA locus with Lp(a) plasma levels in 14,500 participants of six white European cohorts. *Atherosclerosis.***217:** 447–451.
20. Smith, E. N., W. Chen, M. Kahonen, J. Kettunen, T. Lehtimaki, L. Peltonen, O. T. Raitakari, R. M. Salem, N. J. Schork, M. Shaw, et al. 2010. Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa heart study. *PLoS Genet.* **6:** e1001094.
21. Ken-Dror, G., P. J. Talmud, S. E. Humphries, and F. Drenos. 2010. APOE/C1/C4/C2 gene cluster genotypes, haplotypes and lipid levels in prospective coronary heart disease risk among UK healthy men. *Mol. Med.* **16:** 389–399.
22. Marcovina, S. M., M. L. Koschinsky, J. J. Albers, and S. Skarlatos. 2003. Report of the National Heart, Lung, and Blood Institute Workshop on Lipoprotein(a) and Cardiovascular Disease: recent advances and future directions. *Clin. Chem.* **49:** 1785–1796.
23. Lamon-Fava, S., S. M. Marcovina, J. J. Albers, H. Kennedy, C. Deluca, C. C. White, L. A. Cupples, J. R. McNamara, L. J. Seman, V. Bongard, et al. 2011. Lipoprotein(a) levels, apo(a) isoform size, and coronary heart disease risk in the Framingham Offspring Study. *J. Lipid Res.* **52:** 1181–1187.
24. Erqou, S., S. Kaptoge, P. L. Perry, E. Di Angelantonio, A. Thompson, I. R. White, S. M. Marcovina, R. Collins, S. G. Thompson, and J. Danesh. 2009. Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. *JAMA.* **302:** 412–423.
25. Genser, B., K. C. Dias, R. Siekmeier, T. Stojakovic, T. Grammer, and W. Maerz. 2011. Lipoprotein (a) and risk of cardiovascular disease–a systematic review and meta analysis of prospective studies. *Clin. Lab.* **57:** 143–156.
26. Nordestgaard, B. G., M. J. Chapman, K. Ray, J. Boren, F. Andreotti, G. F. Watts, H. Ginsberg, P. Amarenco, A. Catapano, O. S. Descamps, et al. 2010. Lipoprotein(a) as a cardiovascular risk factor: current status. *Eur. Heart J.* **31:** 2844–2853.
27. Kraft, H. G., A. Lingenhel, S. Kochl, F. Hoppichler, F. Kronenberg, A. Abe, V. Muhlberger, D. Schonitzer, and G. Utermann. 1996. Apolipoprotein(a) kringle IV repeat number predicts risk for coronary heart disease. *Arterioscler. Thromb. Vasc. Biol.* **16:** 713–719.
28. Clarke, R., J. F. Peden, J. C. Hopewell, T. Kyriakou, A. Goel, S. C. Heath, S. Parish, S. Barlera, M. G. Franzosi, S. Rust, D. Bennett, et al. 2009. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N. Engl. J. Med.* **361:** 2518–2528.
29. Lanktree, M. B., S. S. Anand, S. Yusuf, and R. A. Hegele. 2010. Comprehensive analysis of genomic variation in the LPA locus and its relationship to plasma lipoprotein(a) in South Asians, Chinese, and European Caucasians. *Circ. Cardiovasc. Genet.* **3:** 39–46.
30. Li, K. M., D. E. Wilcken, and N. P. Dudman. 1994. Effect of serum lipoprotein(a) on estimation of low-density lipoprotein cholesterol by the Friedewald formula. *Clin. Chem.* **40:** 571–573.
31. Kronenberg, F., A. Lingenhel, K. Lhotta, B. Rantner, M. F. Kronenberg, P. Konig, J. Thiery, M. Koch, A. von Eckardstein, and H. Dieplinger. 2004. Lipoprotein(a)- and low-density lipoprotein-derived cholesterol in nephrotic syndrome: impact on lipid-lowering therapy? *Kidney Int.* **66:** 348–354.
32. Scanu, A. M., and J. Hinman. 2002. Issues concerning the monitoring of statin therapy in hypercholesterolemic subjects with high plasma lipoprotein(a) levels. *Lipids.* **37:** 439–444.
33. Ben-Zeev, O., G. Stahnke, G. Liu, R. C. Davis, and M. H. Doolittle. 1994. Lipoprotein lipase and hepatic lipase: the role of asparagine-linked glycosylation in the expression of a functional enzyme. *J. Lipid Res.* **35:** 1511–1523.
34. Mangravite, L. M., M. W. Medina, J. Cui, S. Pressman, J. D. Smith, M. J. Rieder, X. Guo, D. A. Nickerson, J. I. Rotter, and R. M. Krauss. 2010. Combined influence of LDLR and HMGCR sequence variation on lipid-lowering response to simvastatin. *Arterioscler. Thromb. Vasc. Biol.* **30:** 1485–1492.
35. Donnelly, L. A., A. S. Doney, J. Dannfald, A. L. Whitley, C. C. Lang, A. D. Morris, P. T. Donnan, and C. N. Palmer. 2008. A paucimorphic variant in the HMG-CoA reductase gene is associated with lipid-lowering response to statin treatment in diabetes: a GoDARTS study. *Pharmacogenet. Genomics.* **18:** 1021–1026.
36. Iakoubova, O. A., C. H. Tong, C. M. Rowland, T. G. Kirchgessner, B. A. Young, A. R. Arellano, D. Shiffman, M. S. Sabatine, H. Campos, C. J. Packard, et al. 2008. Association of the Trp719Arg polymorphism in kinesin-like protein 6 with myocardial infarction and coronary heart disease in 2 prospective trials: the CARE and WOSCOPS trials. *J. Am. Coll. Cardiol.* **51:** 435–443.
37. Becker, M. L., L. E. Visser, R. H. van Schaik, A. Hofman, A. G. Uitterlinden, and B. H. Stricker. 2009. Common genetic variation in the ABCB1 gene is associated with the cholesterol-lowering effect of simvastatin in males. *Pharmacogenomics.* **10:** 1743–1751.

nature COMMUNICATIONS

ARTICLE

Received 27 Feb 2014 | Accepted 22 Aug 2014 | Published 28 Oct 2014

DOI: 10.1038/ncomms6068 OPEN

Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins

Iris Postmus[1,2,*], Stella Trompet[1,3,*], Harshal A. Deshmukh[4,*], Michael R. Barnes[5,6,*], Xiaohui Li[7], Helen R. Warren[6,8], Daniel I. Chasman[9,10], Kaixin Zhou[4], Benoit J. Arsenault[11], Louise A. Donnelly[4], Kerri L. Wiggins[12], Christy L. Avery[13], Paula Griffin[14], QiPing Feng[15], Kent D. Taylor[7], Guo Li[12], Daniel S. Evans[16], Albert V. Smith[17,18], Catherine E. de Keyser[19,20], Andrew D. Johnson[21], Anton J.M. de Craen[1,2], David J. Stott[22], Brendan M. Buckley[23], Ian Ford[24], Rudi G.J. Westendorp[1,2,25], P. Eline Slagboom[2,26], Naveed Sattar[27], Patricia B. Munroe[6,8], Peter Sever[28], Neil Poulter[28], Alice Stanton[29,30], Denis C. Shields[31,32], Eoin O'Brien[31], Sue Shaw-Hawkins[5,6], Y.-D. Ida Chen[7], Deborah A. Nickerson[33], Joshua D. Smith[33], Marie Pierre Dubé[11], S. Matthijs Boekholdt[34], G. Kees Hovingh[35], John J.P. Kastelein[35], Paul M. McKeigue[36], John Betteridge[37], Andrew Neil[38], Paul N. Durrington[39], Alex Doney[4], Fiona Carr[4], Andrew Morris[4], Mark I. McCarthy[40,41,42], Leif Groop[43], Emma Ahlqvist[43], Welcome Trust Case Control Consortium[†], Joshua C. Bis[12], Kenneth Rice[44], Nicholas L. Smith[45,46,47], Thomas Lumley[12,48], Eric A. Whitsel[13,49], Til Stürmer[13], Eric Boerwinkle[50], Julius S. Ngwa[14], Christopher J. O'Donnell[51,52,53], Ramachandran S. Vasan[54], Wei-Qi Wei[55], Russell A. Wilke[56], Ching-Ti Liu[14], Fangui Sun[14], Xiuqing Guo[7], Susan R. Heckbert[45,46,57], Wendy Post[58], Nona Sotoodehnia[12,59], Alice M. Arnold[44], Jeanette M. Stafford[60], Jingzhong Ding[61], David M. Herrington[62], Stephen B. Kritchevsky[63], Gudny Eiriksdottir[17], Leonore J. Launer[64], Tamara B. Harris[64], Audrey Y. Chu[9], Franco Giulianini[9], Jean G. MacFadyen[9], Bryan J. Barratt[65], Fredrik Nyberg[66,67], Bruno H. Stricker[19,20,68], André G. Uitterlinden[2,19,68], Albert Hofman[2,19], Fernando Rivadeneira[68], Valur Emilsson[17], Oscar H. Franco[19], Paul M. Ridker[9], Vilmundur Gudnason[17,18], Yongmei Liu[61], Joshua C. Denny[55,69], Christie M. Ballantyne[70], Jerome I. Rotter[7], L. Adrienne Cupples[14,51], Bruce M. Psaty[12,45,46,71], Colin N.A. Palmer[4], Jean-Claude Tardif[11], Helen M. Colhoun[4,72], Graham Hitman[73], Ronald M. Krauss[74], J. Wouter Jukema[3,75,76,‡] & Mark J. Caulfield[6,8,‡]

Statins effectively lower LDL cholesterol levels in large studies and the observed interindividual response variability may be partially explained by genetic variation. Here we perform a pharmacogenetic meta-analysis of genome-wide association studies (GWAS) in studies addressing the LDL cholesterol response to statins, including up to 18,596 statin-treated subjects. We validate the most promising signals in a further 22,318 statin recipients and identify two loci, *SORT1/CELSR2/PSRC1* and *SLCO1B1*, not previously identified in GWAS. Moreover, we confirm the previously described associations with *APOE* and *LPA*. Our findings advance the understanding of the pharmacogenetic architecture of statin response.

[1] Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden 2300 RC, The Netherlands. [2] The Netherlands Consortium for Healthy Ageing, Leiden 2300 RC, The Netherlands. [3] Department of Cardiology, Leiden University Medical Center, Leiden 2300 RC, The Netherlands. [4] Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee DD1 9SY, UK. [5] Genome Centre, William Harvey Research Institute, Barts and The London School of Medicine, Queen Mary University of London, London EC1M6BQ, UK. [6] NIHR Barts Cardiovascular Biomedical Research Unit, Queen Mary University of London, London EC1M 6BQ, UK. [7] Institute for Translational Genomics and Population Sciences, Los Angeles BioMedical Research Institute at Harbor-UCLA Medical Center, Torrance, California 90502, USA. [8] Department of Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine, Queen Mary University of London, London EC1M6BQ, UK. [9] Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02215-1204, USA. [10] Harvard Medical School, Boston, Massachusetts 02215, USA. [11] Montreal Heart Institute, Universite de Montreal, Montreal H1T 1C8, Quebec, Canada. [12] Cardiovascular Health Research Unit, Department of Medicine, University of Washington, 98101 Seattle, Washington, USA. [13] Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [14] Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02215, USA. [15] Department of Clinical Pharmacology, Vanderbilt University, Nashville, Tennessee 37240, USA. [16] California Pacific Medical Center Research Institute, San Francisco, California 94107, USA. [17] Icelandic Heart Association, IS-201 Kopavogur, Iceland. [18] University of Iceland, IS-101 Reykjavik, Iceland. [19] Department of Epidemiology, Erasmus Medical Center, 3000 CA Rotterdam, The Netherlands. [20] Health Care Inspectorate, 2595 AN The Hague, The Netherlands. [21] Framingham Heart Study (FHS) of the National Heart, Lung and Blood Institute, Cardiovascular Epidemiology and Human Genomics, Framingham, Massachusetts 01702, USA. [22] Faculty of Medicine, Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow G31 2ER, UK. [23] Department of Pharmacology and Therapeutics, University College Cork, Cork 30, Ireland. [24] Robertson Center for Biostatistics, University of Glasgow, Glasgow G12 8QQ, UK. [25] Leyden Academy of Vitality and Ageing, 2333 AA Leiden, The Netherlands. [26] Department of Molecular Epidemiology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands. [27] Faculty of Medicine, BHF Glasgow Cardiovascular Research Centre, Glasgow G12 8QQ, UK. [28] International Centre for Circulatory Health, Imperial College, London SW7 2AZ, UK. [29] Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin 2, Ireland. [30] Beaumont Hospital, Dublin 9, Ireland. [31] The Conway Institute, University College Dublin, Dublin 4, Ireland. [32] School of Medicine and Medical Sciences, University College Dublin, Dublin 4, Ireland. [33] Department of Genome Sciences, University of Washington, Seattle, Washington 98101, USA. [34] Department of Cardiology, Academic Medical Center, 1100 DD Amsterdam, The Netherlands. [35] Department of Vascular Medicine, Academic Medical Center, 1100 DD Amsterdam, The Netherlands. [36] University of Edinburgh, Edinburgh EH9 3JR, UK. [37] University College, London WC1E 6BT, UK. [38] University of Oxford, Oxford OX1 2JD, UK. [39] Cardiovascular Research Group, School of Biosciences, University of Manchester, Manchester M13 9NT, UK. [40] Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK. [41] Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [42] Oxford NIHR Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK. [43] Department of Clinical Sciences/Diabetes & Endocrinology, Lund University, Malmo 205 02, Sweden. [44] Department of Biostatistics, University of Washington, 98115 Seattle, Washington, USA. [45] Department of Epidemiology, University of Washington, Seattle, Washington 98195, USA. [46] Group Health Research Institute, Group Health Cooperative, Seattle, Washington 98101, USA. [47] Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, Washington 98101, USA. [48] Department of Statistic, University of Auckland, Auckland 1142, New Zealand. [49] Department of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [50] Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. [51] NHLBI Framingham Heart Study, Framingham, Massachusetts 01701, USA. [52] Cardiology Division, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [53] National Heart, Lung and Blood Institute, Bethesda, Maryland 20892, USA. [54] Section of Preventive Medicine and Epidemiology, Department of Medicine, Boston University School of Medicine, and the Framingham Heart Study, Framingham, Massachusetts 01701, USA. [55] Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee 37240, USA. [56] Department of Internal Medicine, Center for IMAGENETICS, Sanford Healthcare, Fargo, North Dakota, 58104 USA. [57] Cardiovascular Health Research Unit, University of Washington, Seattle, Washington 98101, USA. [58] Department of Cardiology, Johns Hopkins University, Baltimore, Maryland 21218, USA. [59] Division of Cardiology, Harborview Medical Center, University of Washington, Seattle 98101, Washington, USA. [60] Division of Public Health Sciences, Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA. [61] Division of Public Health Sciences, Department of Epidemiology and Prevention, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA. [62] Department of Internal Medicine, Section on Cardiology, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA. [63] Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA. [64] Laboratory of Epidemiology, Demography, Biometry, National Institute on Aging, National Institutes of Health, 7201 Wisconsin Avenue, Bethesda, Maryland 20892, USA. [65] Personalised Healthcare and Biomarkers, AstraZeneca, Alderley Park SK10 4TG, UK. [66] AstraZeneca Research and Development, 481 83 Mölndal, Sweden. [67] Unit of Occupational and Environmental Medicine, University of Gothenburg, 405 30 Gothenburg, Sweden. [68] Department of Internal Medicine, Erasmus Medical Center, 3000 CA Rotterdam, The Netherlands. [69] Department of Medicine, Vanderbilt University, Vanderbilt, Tennessee 37240, USA. [70] Department of Medicine, Baylor College of Medicine, Houston, Texas 77030, USA. [71] Department of Health Services, University of Washington, Seattle, Washington 98101, USA. [72] Department of Public Health, University of Dundee, Dundee DD1 9SY, UK. [73] Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK. [74] Children's Hospital Oakland Research Institute, Oakland, California 94609, USA. [75] Durrer Center for Cardiogenetic Research, 1105 AZ Amsterdam, The Netherlands. [76] Interuniversity Cardiology Institute of the Netherlands, 3511 GC Utrecht, The Netherlands. * These authors contributed equally. † Members of Welcome Trust Case Control Consortium are listed at the end of the paper. ‡ These authors jointly supervised the work. Correspondence and requests for materials should be addressed to J.W.L. (email: j.w.jukema@lumc.nl) or M.J.C. to (email: m.j.caulfield@qmul.ac.uk).

The 3-hydroxymethyl-3-methylglutaryl coenzyme A (HMG-CoA) reductase inhibitors, also known as statins, are widely prescribed and are highly effective in the management and prevention of cardiovascular disease. Statin therapy results in a lowering of low-density lipoprotein cholesterol (LDL-C) levels by up to 55%[1] and a 20–30% reduction of cardiovascular events[2]. Despite the clinical efficacy of statins in a wide range of patients[2], interindividual variability exists with regard to LDL-C-lowering response as well as efficacy in reducing major cardiovascular events[3]. The suggestion that some of this variability may be due, in part, to common pharmacogenetic variation is supported by previous studies that have identified genetic variants associated with differential LDL-C response to statin therapy[4–6].

A small number of genome-wide association studies (GWAS) have previously identified loci associated with statin response on a genome-wide level. A GWAS in the JUPITER trial identified three genetic loci, *ABCG2* (rs2199936), *LPA* (rs10455872) and *APOE* (rs7412), that were associated with percentage LDL-C reduction following rosuvastatin therapy[7]. In the CARDS and ASCOT studies, single nucleotide polymorphisms (SNPs) at *LPA* (rs10455872) and *APOE* (rs445925 and rs4420638) were associated with LDL-C response to atorvastatin treatment[8]. A combined GWAS in three statin trials identified a SNP within *CLMN* (rs8014194) that is associated with the magnitude of statin-induced reduction in plasma cholesterol[9]. However, two other GWAS identified no genetic determinants of LDL-C response to statin therapy at a genome-wide significant level[6,10].

On the basis of these studies, as well as previous candidate gene studies[4,6], the only genetic variants that have been consistently identified to be associated with variation in LDL-C response to statin therapy, irrespective of statin formulation, are located at or nearby *APOE* and *LPA*. To determine whether additional loci may influence LDL-C response to statins, we formed the Genomic Investigation of Statin Therapy (GIST) consortium and conducted a pharmacogenetic meta-analysis using GWAS data sets from randomized controlled trials (RCTs) and observational studies. We identify two loci not previously identified in GWAS, *SORT1/CELSR2*/PSRC1 and *SLCO1B1*. In addition, we confirm the associations within the *APOE* and *LPA* genes. These findings will extend the knowledge of the pharmacogenetic architecture of statin response.

Results

First-stage meta-analysis. The GIST consortium includes 6 RCTs ($n=8,421$ statin recipients) and 10 observational studies ($n=10,175$ statin recipients) that participated in the first stage (see Methods; Supplementary Tables 1 and 2; Supplementary Notes 1 and 2). To search for genetic variants associated with differential LDL-C response to statin therapy, each study independently performed a GWAS among statin users, using the difference between the natural log-transformed LDL-C levels on- and off-treatment as the response variable (see Methods).

The first-stage meta-analysis identified three loci, including 13 SNPs, that attained genome-wide significance ($P<5\times10^{-8}$) for association with LDL-C response to statin treatment (Fig. 1; Table 1). The most significant association was for a SNP on chromosome 19, at *APOE* (rs445925, minor allele frequency (MAF) = 0.098, $\beta=-0.043$, s.e. = 0.005, $P=1.58\times10^{-18}$; Fig. 2a), indicating that carriers of the rs445925 SNP respond to statins with an additional 4.3% increase per allele in LDL-C lowering effect compared with non-carriers. The second strongest association was with a SNP at *LPA* on chromosome 6 (rs10455872, MAF = 0.069, $\beta=0.041$, s.e. = 0.006, $P=1.95\times10^{-11}$; Fig. 2b), indicating a 5.9% smaller LDL-C lowering per minor allele for carriers of the SNP compared with non-carriers. Associations at both loci have previously been described[7,8]. A third genome-wide significant association was found with a SNP at *RICTOR* on chromosome 5 (rs13166647, MAF = 0.230, $\beta=-0.253$, s.e. = 0.046, $P=4.50\times10^{-8}$), although genotypes for this SNP were only available in two studies within the first stage ($n=2,144$).

Second-stage meta-analysis. We selected 246 SNPs with $P<5\times10^{-4}$ from 158 loci for further investigation in three additional studies comprising up to 22,318 statin-treated subjects (see Methods; Supplementary Tables 1 and 5; Supplementary Note 3). This second stage confirmed the genome-wide significant associations between variations within the *APOE* and *LPA* loci and LDL-C response, as observed in the first stage (Table 1; Supplementary Fig. 2; Supplementary Table 5). In addition, SNPs at two new loci with P values between 6.70×10^{-7} and 2.26×10^{-6} in the first phase were shown to be significantly associated with statin-induced LDL-C lowering after statin treatment in the total combined meta-analysis at a genome-wide level: *SORT1/CELSR2/PSRC1* (rs646776, $\beta=-0.013$, s.e. = 0.002, $P=1.05\times10^{-9}$ and rs12740374, $\beta=-0.013$, s.e. = 0.002, $P=1.05\times10^{-9}$; Fig 2c) and *SLCO1B1* (rs2900478, $\beta=0.016$, s.e. = 0.003, $P=1.22\times10^{-9}$; Fig 2d), indicating an additional 1.5% increase per allele in LDL-C lowering effect for carriers of the *SORT1/CELSR2/PSRC1* SNP and a 1.6% smaller LDL-C lowering per minor allele for carriers of the *SLCO1B1* SNP.

The six next-ranked SNPs with P values just below 5×10^{-8} in the combined meta-analysis, including the two SNPs at *RICTOR* (rs13166647 and rs13172966), were selected for additional genotyping in the Scandinavian ASCOT participants (see Methods). None of these six SNPs reached genome-wide significance after this additional genotyping (Supplementary Table 6). Therefore, our overall genome-wide significant findings were the SNPs at *APOE*, *LPA*, *SORT1/CELSR2/PSRC1* and *SLCO1B1*.

Subfraction analyses. To extend our results for the novel GWAS finding *SORT1/CELSR2/PSRC1*, we performed additional association analyses, using measurements of cholesterol levels in four LDL subfractions (large, medium, small and very small) from two of the trials in GIST, CAP and PRINCE (Table 2; see Methods). The minor allele of *SORT1* rs646776 was associated with greater statin-induced reductions in levels of all LDL subfractions, and there was a nonsignificant trend for larger effect sizes and greater statistical significance for lowering of small and very small LDL (Table 2). In contrast, the *APOE* SNP associated with greater

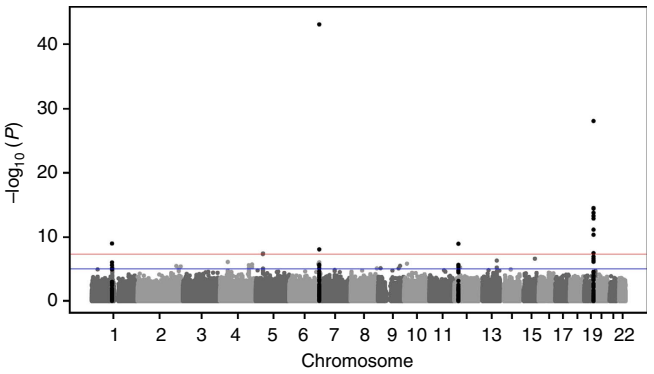


Figure 1 | Results of the GWAS meta-analysis. Manhattan plot presenting the $-\log_{10}P$ values from the combined meta-analysis ($n=40,914$) on LDL-C response after statin treatment. *P* values were generated using linear regression analysis.

Table 1 | Genome-wide significant associations in stage 1, stage 2 and combined meta-analysis.

Chr	Position	Lead SNP	Gene	Coding allele	Noncoding allele	Phase	*N*	Frequency-coding allele	Beta*	s.e.	% Extra reduction†	*P* value
1	109620053	rs646776	*SORT1/ CELSR2/ PSRC1*	C	T	Stage 1	16,697	0.230	−0.015	0.003	1.5	6.70×10^{-7}
						Stage 2	21,902	0.216	−0.010	0.003	1.0	2.43×10^{-4}
						Combined	38,599		−0.013	0.002	1.3	1.05×10^{-9}
6	160930108	rs10455872	*LPA*	G	A	Stage 1	12,981	0.069	0.041	0.006	−4.1	1.95×10^{-11}
						Stage 2	18,075	0.087	0.059	0.005	−5.9	7.14×10^{-35}
						Combined	31,056		0.052	0.004	−5.2	7.41×10^{-44}
12	21260064	rs2900478	*SLCO1B1*	A	T	Stage 1	16,749	0.165	0.016	0.003	−1.6	2.26×10^{-6}
						Stage 2	7,504	0.164	0.017	0.006	−1.7	3.54×10^{-3}
						Combined	24,253		0.016	0.003	−1.6	1.22×10^{-9}
19	50107480	rs445925	*APOE*	A	G	Stage 1	13,909	0.098	−0.043	0.005	4.3	1.58×10^{-18}
						Stage 2	3,613	0.157	−0.088	0.011	8.8	1.41×10^{-15}
						Combined	17,522		−0.051	0.005	5.1	8.52×10^{-29}

Chr, chromosome; SNP, single nucleotide polymorphism.
*Beta for difference between the natural log-transformed on- and off-treatment low-density lipoprotein cholesterol (LDL-C) levels adjusted for natural log-transformed off-treatment LDL-C-, age-, sex- and study-specific covariates. The beta reflects the fraction of differential LDL-C lowering in carriers versus non-carriers of the SNP; a negative beta indicates a better statin response (stronger LDL-C reduction), a positive beta a worse statin response. Betas and *P* values were generated using linear regression analysis.
†This percentage reflects the % extra LDL-C lowering in carriers versus non-carriers of the SNP.

LDL-C response to statins (rs445925) showed a small and non-significant association with change in very small LDL (Table 2). For the minor allele of rs2900478 (*SLCO1B1*), the borderline significant association with smaller magnitude of LDL-C reduction showed a trend for preferential association with larger versus smaller LDL subfractions. The lack of association of rs10455872 (*LPA*) with changes in LDL subfractions is consistent with evidence discussed below that this locus affects levels of lipoprotein(a) (Lp(a)) and not LDL particles. Using generalized estimating equations, we tested the association of log change in each of the LDL subfractions with interactions of the four SNPs. For very small LDL, the association with the rs646776 minor allele was significantly different from that of the other minor alleles ($P=0.03$ after adjustment for multiple testing).

Effects of off-treatment LDL-C. To demonstrate that our findings for LDL-C response to statin treatment are unlikely to be explained through associations with baseline LDL-C levels, we performed a number of additional analyses (see Methods). First, Supplementary Table 7 shows regression coefficients for baseline-adjusted and measurement noise-corrected estimates of the direct effect of genotype on on-treatment LDL-C at the strongest SNPs in the GIST meta-analysis ($P<1\times10^{-8}$), which were available in the CARDS data set. Correcting our effect size estimate further and modelling measurement noise at baseline reduced the apparent effect only slightly for all the markers, suggesting that there is little effect of measurement noise. Next, within the JUPITER trial, additional analyses were performed to determine whether there was an interaction between LDL-C change and statin or placebo allocation. Supplementary Table 8 shows significant *P* values for interaction (all $<5\times10^{-2}$) for SNPs at the four genome-wide significant loci in the GIST meta-analysis, also suggesting that genetic effects on baseline LDL-C as manifested in the placebo group contribute at most only in part to genetic effects on LDL-C response in the statin group.

Genome-Wide Conditional Analysis. To investigate whether there were multiple SNPs within any gene and multiple loci associated with differential LDL-C lowering to statin therapy, we performed a conditional analysis across the genome using the summary statistics of the combined meta-analysis. The results of the Genome-Wide Conditional Analysis (GWCA; see Methods; Supplementary Table 9) showed 14 SNPs independently associated with statin response and these explained ~5% of the variation in LDL-C response to statin treatment. Of the 14 independent SNPs, 6 were genome-wide significant in the combined GWAS meta-analysis (Supplementary Table 5).

Previous findings. In Supplementary Table 10, we performed a look-up in our GWAS meta-analysis for SNPs previously described in the literature (NHGRI Catalogue[11] of Published GWAS and Candidate gene studies) to be associated with statin response, besides the loci associated at a genome-wide level in the current study. None of these SNPs was associated with statin response in our GWAS after correcting for multiple testing.

Functional analyses. Functional characterization of the 246 SNPs selected for the second stage was performed using a range of bioinformatics tools (see Methods). A total of 420 expression quantitative trait loci (eQTL) associations were identified across a wide range of tissues (Supplementary Data 1), which comprised 67 independent gene eQTL associations. Eleven genes, including *APOE*, *SORT1*, *CELSR2* and *PSRC1*, showed eQTLs in liver, which considering its primary role in mediating statin-induced LDL reduction may be particularly relevant to statin response. Putative gene eQTLs were combined with genes annotated to variants in linkage disequilibrium (LD) with LDL-C response-associated variants, resulting in a list of 185 candidate gene loci, defined by 2,681 SNPs (Supplementary Data 2 and 3). To identify statin responsive genes among the candidate loci, gene expression data measured in response to statin treatment in a range of cell lines was retrieved from the Connectivity Map resource[12] (see Methods). Five genes (*APOE, BRCA1, GRPEL1, ADRB2* and *ETV1*) showed convincing evidence of statin responsiveness on the basis of greater than twofold differential expression in response to statin treatment. Eight genes showed suggestive evidence (1.5- to 2-fold change; *TOMM40*, *SREBP1*, *PSRC1*, *BCL3*, *BCAM*, *ANK3*, *SIVA1* and *RANBP9*; Supplementary Data 3).

Finally, involvement in statin response was investigated at a pathway level using GeneGo Metacore (Thomson Reuters[13]). Briefly, 87 literature-reported genes linked to statin response were combined with the 185 candidate gene loci reported here

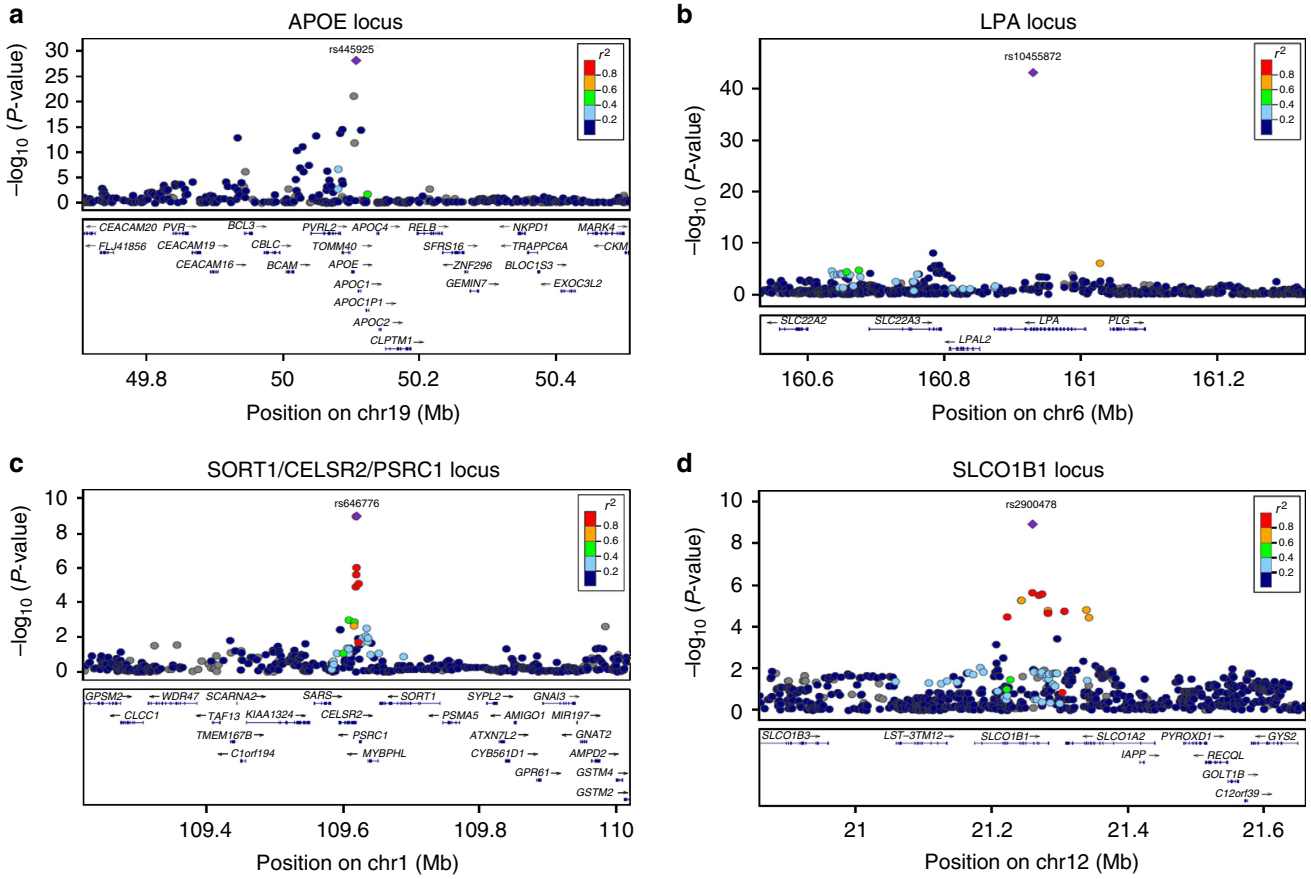


Figure 2 | Regional association plots of the genome-wide significant associations with LDL-C response after statin treatment. The plots show the genome-wide significant associated loci in the combined meta-analysis ($n = 40,914$), the *APOE* locus (**a**), the *LPA* locus (**b**), the *SORT1/CELSR2/PSRC1* locus (**c**) and the *SLCO1B1* locus (**d**) (generated using LocusZoom (http://genome.sph.umich.edu/wiki/LocusZoom)). The colour of the SNPs is based on the LD with the lead SNP (shown in purple). The RefSeq genes in the region are shown in the lower panel. *P* values were generated using linear regression analysis.

Table 2 | Associations of the minor alleles of rs646776, rs445925, rs2900478 and rs10455872 with changes in LDL-C and LDL subfractions in response to statin in the combined CAP and PRINCE studies.

Change*	*SORT1/CELSR2/PSRC1* rs646776 (MAF 0.2)			*APOE* rs445925 (MAF 0.086)			*SLCO1B1* rs2900478 (MAF 0.16)			*LPA* rs10455872 (MAF 0.056)		
	Beta	**s.e.**	***P* value**	**Beta**	**s.e.**	***P* value**	**Beta**	**s.e.**	***P* value**	**Beta**	**s.e.**	***P* value**
LDL-C total	−0.023	0.008	0.003	−0.046	0.018	0.008	0.010	0.005	0.04	0.032	0.019	0.09
Large LDL-C	−0.028	0.014	0.042	−0.075	0.029	0.009	0.02	0.008	0.01	0.036	0.031	0.23
Medium LDL-C	−0.027	0.015	0.075	−0.079	0.032	0.012	0.016	0.009	0.07	0.010	0.034	0.77
Small LDL-C	−0.047	0.018	0.009	−0.071	0.037	0.050	0.002	0.010	0.83	−0.024	0.039	0.54
Very small LDL-C	−0.034	0.009	0.00006	−0.022	0.017	0.202	0.001	0.005	0.90	0.008	0.019	0.67

LDL-C, low-density lipoprotein cholesterol; MAF, minor allele frequency.
*Change: ln (on treatment) − ln (baseline) models adjusted for log (baseline variable), age, sex, body mass index, smoking(y/n) and study (CAP versus PRINCE). Betas and *P* values were assessed using a generalized estimating equation method.

(Supplementary Data 3). A conservative network of direct interactions was constructed between query genes (Supplementary Data 4). The network included 24 genes located in the LDL-C-associated loci (Supplementary Fig. 4). Collectively, our functional and pathway analysis confirms a strong biological and functional role in statin response for several strongly associated gene loci, including *APOE/TOMM40/PVRL2* and *SORT1/CELSR2/PSRC2*.

Discussion

We have performed a meta-analysis of GWAS including more than 40,000 subjects, investigating genetic variants associated with variation in LDL-C lowering on statin treatment independent from associations with baseline LDL-C. We identified four loci at genome-wide significance, including the previously identified *APOE* and *LPA*, and the novel GWAS loci *SORT1/CELSR2/PSRC1* and *SLCO1B1*.

Nine SNPs in the *APOE* gene region reached genome-wide significance for LDL-C response. The minor allele of the lead SNP rs445925, which is a proxy for the apoE ε2 protein variant defining SNP rs7412 (ref. 14), was associated with a larger LDL-C-lowering response to statins compared with carriers of the major allele. The magnitude and direction of the effect size was similar to previously reported findings for the rs445925 variant in

the GWAS study performed in CARDS and ASCOT[8] and of the SNP rs7412 in JUPITER[7]. Since the apoE ϵ2 protein results in increased hepatic cholesterol synthesis, it may also predispose to stronger inhibition of cholesterol synthesis by statin treatment[8,10].

Three independent SNPs at *LPA* were significantly associated with LDL-C response to statins. The minor G allele of the lead SNP rs10455872 was associated with smaller LDL-C reduction than the major allele. This result was similar to the previous GWAS findings for this SNP in the JUPITER trial and the combined ASCOT and CARDS study[7,8]. The rs10455872 SNP was strongly associated with the KIV-2 copy number variant in Lp(a), which encodes variability in apo(a) size and is responsible for ~30% of variance in Lp(a) levels[8,15]. Furthermore, rs10455872 was shown to be strongly associated with plasma Lp(a) levels[16]. Standard assays of LDL-C, as well as the Friedewald formula, include cholesterol that resides in Lp(a)[6,8]. Carriers of this *LPA* variant are characterized by higher Lp(a) levels and a larger proportion of their measured LDL-C resides in Lp(a) particles[8,10]. Since statin therapy does not reduce the number of Lp(a) particles[17], their presence attenuates the measured LDL-C response to statins.

Two SNPs at *SORT1/CELSR2/PSRC1* (rs646776 and rs12740374) on chromosome 1p were associated with an enhanced statin LDL-C response. A similar association was previously observed in a large candidate gene study in HPS[6]; however, we demonstrate this finding now first at a genome-wide significance level. The minor allele of rs12740374 has been shown to generate a binding site for the transcription factor C/EBPa[18]. Transcription results in upregulation of hepatic expression of three genes at this locus, *SORT1*, *CELSR2* and *PSRC1* (ref. 18), which we also showed in our eQTL analysis (Supplementary Data 1). Of these, *SORT1* is most notable, in that it encodes the multifunctional intracellular trafficking protein sortilin, which has been shown to bind tightly to apoB[19]. Sortilin-induced lowering of plasma LDL-C results from two mechanisms: reduced secretion of apoB-containing precursors, and, perhaps of greater importance, increased hepatic LDL uptake via binding to sortilin at the cell surface, with subsequent internalization and lysosomal degradation[19]. Notably, the minor allele of rs646776 is preferentially associated with lower levels of small and very small LDL (Table 2), suggesting that sortilin is of particular importance for regulating levels of these particles[18]. Smaller LDL subfractions have been shown to be relatively enriched in particles with reduced LDL receptor binding affinity and cellular uptake[20], a property that may contribute to their associations with increased risk for cardiovascular disease[21]. This property may also underlie the diminished efficacy of statins for reduction of these particles (Supplementary Fig. 3)[22], since statins act to reduce LDL-C levels to a large extent by increasing LDL receptor expression as a result of upregulation of the transcription factor SREBP2, whereas *SORT1* is not regulated by this mechanism. Hence, the greater statin-mediated reduction of LDL-C among carriers of the rs646776 minor allele could be attributed to relative depletion of LDL particles dependent on sortilin for clearance and hence a residually greater proportion of those LDL particles whose uptake is more dependent on the LDL receptor than on sortilin.

Notably, the strong association of rs646776 with statin-induced reductions in small and very small LDL particles contrasts to the weaker associations of changes in these particles with rs445925, likely the result of differing mechanisms underlying the effects of these SNPs on statin response. As noted above, rs445925 is a proxy for the SNP defining the apoE ϵ2 protein variant that is thought to predispose to heightened statin response as a result of greater statin inhibition of cholesterol synthesis and hence upregulation of SREBP and LDL receptor activity.

The *SLCO1B1* rs2900478 minor allele was associated with a smaller LDL-C reduction in response to statin treatment. *SLCO1B1* encodes the organic anion-transporting polypeptide OATP1B1 and facilitates the hepatic uptake of statins[23]. SNP rs2900478 is in strong LD ($r^2 = 0.89$) with rs4149056, which represents the Val174Ala substitution resulting in complete loss of function. In the HPS trial, which used simvastatin, this candidate gene SNP was associated with a 1% lower LDL-C reduction per allele[6]. Single-dose studies have shown that the observed area under the curve of plasma level of active simvastatin after a dose of 40 mg was 221% higher in rs4149056 CC homozygotes compared with rs4149056 TT homozygotes, as compared with atorvastatin 20 mg (144% higher for CC versus TT) and rosuvastatin 40 mg (117% higher for CC versus TT)[24]. This finding results from the slower hepatic uptake of statins caused by the genetic variant, which would also be expected to result in a reduction in the cholesterol-lowering effect[25]. In a GWAS of the genetic risk factors for simvastatin-induced myopathy, *SLCO1B1* showed the strongest association[25]. Homozygous carriers of the *SLCO1B1* variant had a 16.9 times higher risk for myopathy compared with non-carriers. This might have led to a decrease in study medication adherence, and consequently a decreased effect on LDL-C in carriers of this SNP. In addition, previous analysis in the GoDARTS study showed that the effect of the *SLCO1B1* gene on statin efficacy was abolished after removal of individuals who showed signs of intolerance[26].

GWCA identified three independent loci in the *APOE* gene region and two loci in the *LPA* gene region (Supplementary Table 9). GWCA also showed several other loci with $P < 5 \times 10^{-8}$ that were not GWAS significant on single-SNP analysis (*HGD*, *RNF175*, *ISCA1L-HTR1A*, *GLIS3-SLC1A1*, *LOC100128657*, *NKX2-3-SLC25A28* and *PELI2*). These findings will require replication in independent, larger data sets. The significant SNPs in the GWCA analysis explained ~5% of the variation in LDL-C response to statin treatment. Whether this 5% is clinically relevant should be investigated by other studies. For example, it would be of interest to investigate whether this differential LDL-C lowering is also associated with differential event reduction by statin treatment.

In the current study, we combined the results of 6 randomized clinical trials and 10 observational studies in the first stage. This approach resulted also in combining several types of statins, since different statins were studied in the trials and within the observational studies (Supplementary Table 2). This, and the variation in statin dosage during follow-up for an individual, is a limitation of the current study, since, for example, the impact of the *SLCO1B1* variant on statin pharmacogenetics is known to be highly dependent on statin type and dose[24,27]. To overcome this limitation, the individual study analyses were adjusted for statin dose. Dividing the actual statin dose given by the statin-specific dose equivalent (Supplementary Table 3) gives the statin-adjusted equivalent based on the daily dosages required to achieve a mean 30% LDL-C reduction. Using this table, we made the different statin dosages and types comparable within the studies. To correct for between-study variance, we used a fixed effect meta-analysis with inverse variance weighting. Since we observed that the *SLCO1B1* gene was genome-wide significantly associated with LDL lowering, this highlights the thoroughness of our analytical approach, in which the analyses were correctly adjusted for the type and dose of statins used (Supplementary Table 3). Moreover, a comparison of the estimates of the SNPs between the RCTs (where there are no intra-individual differences in dosages) with the estimates of the SNPs in the observational studies showed large homogeneity between the estimates in the various study designs (Supplementary Fig. 2), indicating that our adjustment for dosage seems to be sufficient within this study.

Another possible limitation of the current study is the influence of the identified genetic variants on baseline LDL-C levels. In pharmacogenetic studies investigating the LDL-C-lowering response to statins, it is important to eliminate the effect of association between the genetic variant and baseline LDL-C levels, since those findings may confound the response to treatment associations. Previous large GWAS studies have shown strong associations between baseline LDL-C levels and genetic variants in *SORT1/CELSR2/PSRC1*, *APOE* and *LPA*[28]. To eliminate those possible confounding effects, our response to treatment analyses were adjusted for baseline LDL-C levels. In addition, additional analysis in CARDS and JUPITER suggests no or little influence of genetic associations with baseline LDL-C on the genetic effects on LDL-C-lowering response.

In conclusion, this study is the largest meta-analysis of GWAS for LDL-C response to statin therapy conducted to date. Our results demonstrate that apart from the previously identified *APOE* and *LPA* loci, two new loci, *SORT1/CELSR2/PSRC1* and *SLCO1B1*, also have a modest but genome-wide significant effect on LDL-C response. The minor alleles of the *APOE* rs445925 and *SORT1/CELSR2/PSRC1* rs646776 SNPs were associated with a larger statin response, whereas the minor alleles of the *LPA* rs10455872 and *SLCO1B1* rs2900478 SNPs were associated with a smaller statin response. Our findings advance the understanding of the pharmacogenetic architecture of statin response.

Methods

Study populations. The meta-analysis was conducted in the GIST consortium, which includes data from 8 randomized controlled statin trials (RCTs) and 11 prospective, population-based studies. The initial analysis (first stage) was performed in 8,421 statin-treated subjects from 6 RCTs (ASCOT, CARDS, CAP, PRINCE, PROSPER and TNT) and 10,175 statin-treated subjects from 10 observational studies (AGES, ARIC, BioVU, CHS, FHS, GoDARTS I, GoDARTS II, Health ABC, HVH and MESA). Further investigation (second stage) was performed in 21,975 statin-treated subjects from two randomized trials (HPS and JUPITER) and one observational study (Rotterdam Study). Six SNPs were additionally genotyped in the Scandinavian participants of the ASCOT study. The details of the first- and second-stage studies can be found in the Supplementary Tables 1 and 2 and Supplementary Notes 1 and 2.

Subjects. Response to statin treatment was studied in statin-treated subjects only and not in those treated with placebo. Subjects included in the observational studies' analysis should be treated with statins and have LDL-C measurements before and after start of statin treatment. Subjects of reported or suspected non-European ancestry were excluded. All participants gave written informed consent and the study was approved by all institutional ethics committees.

Outcome measurements. The response to statin treatment was defined as the difference between the natural log-transformed on- and off-treatment LDL-C levels. The beta of the corresponding regression thus reflects the fraction of differential LDL lowering in carriers versus non-carriers of the SNP. For observational studies, the on-treatment LDL-C levels were taken into account for all kinds of prescribed statins, at any dosage, for any indication and for at least 4 weeks before measurement. Characteristics of on- and off-treatment LDL-C levels and statins used in each study are shown in Supplementary Table 2. For each individual, at least one off-treatment LDL-C measurement and at least one on-treatment LDL-C measurement were required. When multiple on- or off-treatment measurements were available, the mean of the cholesterol measurements was used. Subjects with missing on- or off-treatment measurements were excluded, with the exception of the GoDARTS cohorts for which missing off-treatment LDL-C levels were estimated using imputation methods (Supplementary Note 2). In the HPS, proportional LDL-C response was defined by the changes in natural log lipid levels from the screening visit before starting statin therapy to the randomization visit[6].

Genotyping and imputation. Genotyping, quality control, data cleaning and imputation were performed independently in each study using different genetic platforms and software as outlined in Supplementary Table 4. In all studies, genotyping was performed using Illumina, Affymetrix or Perlegen genotyping arrays, and MACH, Impute or BIMBAM software was used for imputation.

GWAS analysis. Each study independently performed the GWAS on the difference between natural log-transformed on- and off-treatment LDL-C levels. To control for possible associations with off-treatment LDL-C levels, analyses were adjusted for the natural log-transformed off-treatment LDL-C level. An additive genetic model was assumed and tested using a linear regression model. For imputed SNPs, regression analysis was performed onto expected allele dosage. Analyses were additionally adjusted for age-, sex- and study-specific covariates (for example, ancestry principal components or country). Analyses in the observational studies were, if available, additionally adjusted for the statin dose by the natural logarithm of the dose equivalent as defined in Supplementary Table 3. This table shows the dose equivalent per statin type; dividing the statin dosage of an individual by the dose equivalent shown in Supplementary Table 3 will give the adjusted statin dosage.

Quality control and meta-analysis. Centrally, within each study, SNPs with MAF $<1\%$ or imputation quality <0.3 were excluded from the analysis. QQ-plots were assessed for each study to identify between-study differences (Supplementary Fig. 1). The software package METAL was used for performing the meta-analysis (http://www.sph.umich.edu/csg/abecasis/Metal/index.html). A fixed effects, inverse variance weighted approach was used. Using an inverse variance weighted meta-analysis will give smaller weights to studies with large s.e.. To correct for possible population stratification, genomic control was performed by adjusting the within-study findings and the meta-analysis results for the genomic inflation factor.

Second stage. SNPs with P values $<5\times10^{-4}$ in the first-stage meta-analysis were selected for further investigation in a second stage. A maximum of two SNPs per locus were selected, based on statistical significance, except for the *APOE* locus, for which all genome-wide significant associated SNPs were selected for validation. A total of 246 SNPs, within 158 independent loci, were selected for the second stage, which was performed in the JUPITER trial, HPS study and the Rotterdam Study, which all had GWAS data and response to statin treatment available. For 2 of the 246 SNPs, a proxy was used in the JUPITER trial, and 31 SNPs were not available, nor was a proxy SNP. HPS provided data on 151 directly genotyped SNPs from GWAS and IPLEX experiments, including 48 of the requested SNPs and 103 proxy SNPs ($r^2>0.8$). Analysis in HPS was not adjusted for ln baseline LDL-C levels. In addition, the number of subjects with data varied from SNP-to-SNP and ranges from $\sim$4,000 for variants with GWAS data to $\sim$18,000 for some candidate genes. Results of the first and second stage were combined using fixed effects, inverse variance weighted meta-analysis and analysed by METAL. As a third stage, six SNPs with P values $5\times10^{-8}<P<5\times10^{-7}$ in the combined meta-analysis were selected for additional genotyping in the Scandinavian participants of the ASCOT study. Kaspar assays were designed for four of the SNPs using the KBioscience Primerpicker software, and oligos were provided by Intergrated DNA technologies (http://eu.idtdna.com/site). Full Kaspar methodology is available from LGC SNP genotyping (http://www.lgcgenomics.com/genotyping/kasp-genotyping-reagents/). Two SNPs (rs981844 and rs13166647) were genotyped using Taqman assays supplied by Life Technologies (http://www.lifetechnologies.com/uk/en/home.html) using the standard Taqman protocol. Results of the additional genotyping were combined with results from the first and second stages using a fixed effects, inverse variance weighted meta-analysis and analysed by METAL.

Determination of changes in LDL subfractions. LDL subclasses were analysed as described previously[29] using non-denaturing gradient gel electrophoresis of fasting plasma samples taken at baseline and after 6 weeks of simvastatin 40 mg per day (CAP study, $n=579$) or 12 weeks of pravastatin 40 mg per day (PRINCE study, $n=1,284$). Aliquots of 3.0 ml of whole plasma were mixed 1:1 with a sampling buffer of 20% sucrose and 0.25% bromophenol blue. Electrophoresis of samples and size calibration standards was performed using 2–14% polyacrylamide gradients at 150 V for 3 h following a 15-min pre-run at 75 V. Gels were stained with 0.07% Sudan black for 1 h and stored in a 0.81% acetic acid, 4% methanol solution until they were scanned by computer-assisted densitometry for determination of areas of LDL IVb (22.0–23.2 nm), LDL IVa (23.3–24.1 nm), LDL IIIb (24.2–24.6 nm), LDL IIIa (24.7–25.5 nm), LDL IIb (25.6–26.4 nm), LDL IIa (26.5–27.1 nm) and LDL I (27.2–28.5 nm). The cholesterol concentrations of the subfractions (mg dl^{-1} plasma) were determined by multiplying percent of the total stained LDL area for each subfraction by the LDL-C for that sample. For genetic association analyses, subfractions were grouped into large LDL (LDL I + IIa), medium LDL (LDL IIb), small LDL (LDL IIIa) and very small LDL (LDL IIIb + IVa + IVb) as described previously[18]. A generalized estimating equation method was used to test the association of log change with the interaction of the four SNPs by LDL subfraction.

Effect of off-treatment LDL-C. Effects of genetic variation on treatment response as measured by on-treatment LDL-C could be mediated through effects on the off-treatment LDL-C. To evaluate whether genetic on-treatment LDL-C likely reflects residual effect on off-treatment LDL-C, it is necessary to adjust for the off-treatment LDL-C levels and to correct the maximum likelihood estimate of the adjusted effect of genotype on on-treatment value for the noise in off-treatment values (the noise is both random measurement error and intra-individual variation in usual LDL-C). This analysis was only carried out in CARDS in which multiple baseline measurements were available. From the rules of path analysis, we calculated the direct effect γ of genotype on an on-treatment trait value as $\beta-\alpha\delta(1-\rho)/\rho$, where

β is the coefficient of regression for on-treatment trait value on genotype adjusted for measured off-treatment value, α is the coefficient of regression of baseline LDL on genotype, ρ is the intraclass correlation between replicate measurements of off-treatment values and δ is the coefficient of regression for on-treatment value on observed off-treatment value[8]. For these calculations, we used $\rho = 0.8$ as a plausible value for the intraclass correlation based on the within-person correlation in LDL-C values taken over two off-treatment visits in CARDS. The interaction of candidate SNPs with statin versus placebo allocation was assessed in the JUPITER trial, since this study was not involved in the first-stage meta-analysis. Regression models were applied to the combined population of statin- and placebo-treated subjects by including extra terms encoding placebo allocation and the product of placebo allocation with SNP minor allele dose[7].

GWCA using Genome-Complex Trait Analysis. There may be multiple causal variants in a gene and the total variation that could be explained at a locus may be underestimated if only the most significant SNP in the region is selected. To identify independent SNPs, we ideally can perform a conditional analysis, starting with the top associated SNP, across the whole genome followed by a stepwise procedure of selecting additional SNPs, one by one, according to their conditional *P* values. Such a strategy would allow the discovery of more than two associated SNPs at a locus. To identify independent SNPs across the genome-wide data, we used an approximate conditional and joint analysis approach implemented in Genome-Complex Trait Analysis (GCTA) software (http://www.complex-traitgenomics.com/software/gcta/). We used summary-level statistics from the first- and second-stage-combined meta-analysis and LD corrections between SNPs estimated from CARDS GWAS data. SNPs on different chromosomes or more than 10 Mb distant are assumed to be in linkage equilibrium. The model selection process in GCTA starts with the most significant SNP in the single-SNP meta-analysis across the whole genome with *P* value $< 5 \times 10^{-7}$. In the next step, it calculates the *P* values of all the remaining SNPs conditional on the top SNP that have already been selected in the model. To avoid problems due to colinearity, if the squared multiple correlations between a SNP to be tested and the selected SNP(s) is larger than a cut-off value, such as 0.9, the conditional *P* value for that SNP will be set to 1. Select the SNPs with minimum conditional *P* value that is lower than the cut-off *P* value. Fit all the selected SNPs jointly in a model and drop the SNPs with the *P* value that is greater than the cut-off *P* value. This process is repeated until no SNPs can be added or removed from the model.

Pathway analysis and construction of a statin response network. Genes showing evidence of association (based on direct association or LD (HapMap CEU $r^2 > 0.8$)) were reviewed for evidence of involvement in statin response at a pathway level using GeneGo Metacore (Thomson Reuters (portal.genego.com)). A statin response network was constructed in two stages. First, all genes with a literature-reported involvement in statin response (based on Medical Subject Headings (MeSH)) were identified using GeneGo MetaCore (Supplementary Data 3). Second, these genes were combined with all genes in associated loci (including genes in LD) and a network was constructed based on direct interactions only. By including direct interactions only, we created a conservative network of direct gene interactions that have been consistently linked to statin response in the literature.

eQTL analysis. LDL-C-associated index SNPs (246 SNPs) were used to identify 1,443 LD proxy SNPs displaying complete LD ($r^2 = 1$) across four HapMap builds in European ancestry samples (CEU) using the SNAP tool (http://www.broadinstitute.org/mpg/snap/). The primary index SNPs and LD proxies were searched against a collected database of expression SNP (eSNP) results, including the following tissues: fresh lymphocytes[30], fresh leukocytes[31], leukocyte samples in individuals with Celiac disease[32], whole-blood samples[33–36], lymphoblastoid cell lines (LCL) derived from asthmatic children[37,38], HapMap LCL from three populations[39], a separate study on HapMap CEU LCL[40], additional LCL population samples[41–43] (Mangravite *et al.*, unpublished), CD19+ B cells[44], primary phytohaemagglutinin-stimulated T cells[41], CD4+ T cells[45], peripheral blood monocytes[44,46,47], CD11+ dendritic cells before and after *Mycobacterium* tuberculosis infection[48], omental and subcutaneous adipose[33,43,49], stomach[49], endometrial carcinomas[50], ER+ and ER− breast cancer tumour cells[51], brain cortex[46,52,53], prefrontal cortex[54,55], frontal cortex[56], temporal cortex[53,56], pons[56], cerebellum[53,56], three additional large studies of brain regions including prefrontal cortex, visual cortex and cerebellum, respectively[57], liver[49,58,59], osteoblasts[60], ileum[49,61], lung[62], skin[43,63] and primary fibroblasts[41]. Micro-RNA QTLs were also queried for LCL[64] and gluteal and abdominal adipose[65]. The collected eSNP results met the criteria for association with gene expression levels as defined in the original papers. In each case where a LDL-C-associated SNP or proxy was associated with a transcript, we further examined the strongest eSNP for that transcript within that data set (best eSNP), and the LD between the best eSNP and GIST-selected eSNPs to estimate the concordance of the LDL-C and expression signals.

Statin response connectivity map analysis. The Connectivity Map (Cmap) data set is available at the Broad Institute (www.broadinstitute.org/cmap) and contains more than 7,000 expression profiles representing 1,309 compounds used on five different cultured human cancer cell lines (MCF7, ssMCF7, HL60, PC3 and SKMEL5). We selected (prostate tumour-derived) PC3 cells as they showed the most responsiveness to statins at a genome-wide level. Four statins were included in our analysis, including pravastatin, atorvastatin, simvastatin and rosuvastatin. PC3 Instance reference files for each statin treatment were extracted (as defined by Lamb *et al.*[12]), that is, a treatment associated to its control pair. Transcripts were considered to show evidence of differential expression with a fold change > 2. A fold change > 1.5 was considered to be suggestive of differential expression only.

Exploration of functional impact among directly and indirectly associated variants. Genes and variants across all LDL-C-associated loci were investigated for evidence of functional perturbation using a range of bioinformatics tools and databases. Variants showing LD (CEU $r^2 > 0.8$) with associated variants were explored for impact on coding gene function using Annovar[66] and regulatory function using a combination of HaploReg[67] and Regulomedb[68], which both draw on comprehensive data from the Encyclopedia of DNA Elements (ENCODE)[69] and the NIH Roadmap Epigenomics consortium[70]. Building on the functional annotation, we also identified variants that were shown to mediate eQTLs. Genes in associated loci were also used to query the NIH connectivity map for evidence of differential expression in PC3 cell lines treated with pravastatin, simvastatin and rosuvastatin. By combining a wide range of functional data and pathway support, we were able to build up a view of genes with the highest level of support in statin response.

References

1. Davidson, M. H. & Toth, P. P. Comparative effects of lipid-lowering therapies. *Prog. Cardiovasc. Dis.* **47,** 73–104 (2004).
2. Baigent, C. *et al.* Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. *Lancet* **376,** 1670–1681 (2010).
3. Mangravite, L. M., Thorn, C. F. & Krauss, R. M. Clinical implications of pharmacogenomics of statin treatment. *Pharmacogenomics J.* **6,** 360–374 (2006).
4. Postmus, I. *et al.* Pharmacogenetics of statins: achievements, whole-genome analyses and future perspectives. *Pharmacogenomics* **13,** 831–840 (2012).
5. Chasman, D. I. *et al.* Pharmacogenetic study of statin therapy and cholesterol reduction. *JAMA* **291,** 2821–2827 (2004).
6. Hopewell, J. C. *et al.* Impact of common genetic variation on response to simvastatin therapy among 18 705 participants in the Heart Protection Study. *Eur. Heart J.* **34,** 982–992 (2013).
7. Chasman, D. I. *et al.* Genetic determinants of statin-induced low-density lipoprotein cholesterol reduction: the Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER) trial. *Circ. Cardiovasc. Genet.* **5,** 257–264 (2012).
8. Deshmukh, H. A. *et al.* Genome-wide association study of genetic determinants of LDL-c response to atorvastatin therapy: importance of Lp(a). *J. Lipid Res.* **53,** 1000–1011 (2012).
9. Barber, M. J. *et al.* Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS ONE* **5,** e9763 (2010).
10. Thompson, J. F. *et al.* Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort. *Circ. Cardiovasc. Genet.* **2,** 173–181 (2009).
11. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
12. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935 (2006).
13. Brennan, R. J., Nikolskya, T. & Bureeva, S. Network and pathway analysis of compound-protein interactions. *Methods Mol. Biol.* **575,** 225–247 (2009).
14. Ken-Dror, G., Talmud, P. J., Humphries, S. E. & Drenos, F. APOE/C1/C4/C2 gene cluster genotypes, haplotypes and lipid levels in prospective coronary heart disease risk among UK healthy men. *Mol. Med.* **16,** 389–399 (2010).
15. Lanktree, M. B., Anand, S. S., Yusuf, S. & Hegele, R. A. Comprehensive analysis of genomic variation in the LPA locus and its relationship to plasma lipoprotein(a) in South Asians, Chinese, and European Caucasians. *Circ. Cardiovasc. Genet.* **3,** 39–46 (2010).
16. Qi, Q., Workalemahu, T., Zhang, C., Hu, F. B. & Qi, L. Genetic variants, plasma lipoprotein(a) levels, and risk of cardiovascular morbidity and mortality among two prospective cohorts of type 2 diabetes. *Eur. Heart J.* **33,** 325–334 (2012).
17. Scanu, A. M. & Hinman, J. Issues concerning the monitoring of statin therapy in hypercholesterolemic subjects with high plasma lipoprotein(a) levels. *Lipids* **37,** 439–444 (2002).
18. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466,** 714–719 (2010).
19. Strong, A. *et al.* Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism. *J. Clin. Invest.* **122,** 2807–2816 (2012).
20. Campos, H., Arnold, K. S., Balestra, M. E., Innerarity, T. L. & Krauss, R. M. Differences in receptor binding of LDL subfractions. *Arterioscler. Thromb. Vasc. Biol.* **16,** 794–801 (1996).

21. Williams, P. T., Zhao, X. Q., Marcovina, S. M., Brown, B. G. & Krauss, R. M. Levels of cholesterol in small LDL particles predict atherosclerosis progression and incident CHD in the HDL-Atherosclerosis Treatment Study (HATS). *PLoS ONE* **8,** e56782 (2013).
22. Choi, Y. J. *et al.* Effects of the PPAR-delta agonist MBX-8025 on atherogenic dyslipidemia. *Atherosclerosis* **220,** 470–476 (2012).
23. Konig, J., Seithel, A., Gradhand, U. & Fromm, M. F. Pharmacogenomics of human OATP transporters. *Naunyn Schmiedebergs Arch. Pharmacol.* **372,** 432–443 (2006).
24. Wilke, R. A. *et al.* The clinical pharmacogenomics implementation consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy. *Clin. Pharmacol. Ther.* **92,** 112–117 (2012).
25. Link, E. *et al.* SLCO1B1 variants and statin-induced myopathy-a genomewide study. *New Engl. J. Med.* **359,** 789–799 (2008).
26. Donnelly, L. A. *et al.* Common nonsynonymous substitutions in SLCO1B1 predispose to statin intolerance in routinely treated individuals with type 2 diabetes: a go-DARTS study. *Clin. Pharmacol. Ther.* **89,** 210–216 (2011).
27. Niemi, M., Pasanen, M. K. & Neuvonen, P. J. Organic anion transporting polypeptide 1B1: a genetically polymorphic transporter of major importance for hepatic drug uptake. *Pharmacol. Rev.* **63,** 157–181 (2011).
28. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
29. Dreon, D. M., Fernstrom, H. A., Williams, P. T. & Krauss, R. M. Reduced LDL particle size in children consuming a very-low-fat diet is related to parental LDL-subclass patterns. *Am. J. Clin. Nutr.* **71,** 1611–1616 (2000).
30. Goring, H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* **39,** 1208–1216 (2007).
31. Idaghdour, Y. *et al.* Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat. Genet.* **42,** 62–67 (2010).
32. Heap, G. A. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* **2,** 1 (2009).
33. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
34. Fehrmann, R. S. *et al.* Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7,** e1002197 (2011).
35. Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* **21,** 48–54 (2013).
36. Sasayama, D. *et al.* Identification of single nucleotide polymorphisms regulating peripheral blood mRNA expression with genome-wide significance: an eQTL study in the Japanese population. *PLoS ONE* **8,** e54967 (2013).
37. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39,** 1202–1207 (2007).
38. Liang, L. *et al.* A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.* **23,** 716–726 (2013).
39. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39,** 1217–1224 (2007).
40. Kwan, T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40,** 225–231 (2008).
41. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325,** 1246–1250 (2009).
42. Cusanovich, D. A. *et al.* The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum. Mol. Genet.* **21,** 2111–2123 (2012).
43. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44,** 1084–1089 (2012).
44. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44,** 502–510 (2012).
45. Murphy, A. *et al.* Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum. Mol. Genet.* **19,** 4745–4757 (2010).
46. Heinzen, E. L. *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.* **6,** e1 (2008).
47. Zeller, T. *et al.* Genetics and beyond-the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5,** e10693 (2010).
48. Barreiro, L. B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl Acad. Sci. USA* **109,** 1204–1209 (2012).
49. Greenawalt, D. M. *et al.* A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.* **21,** 1008–1016 (2011).
50. Kompass, K. S. & Witte, J. S. Co-regulatory expression quantitative trait loci mapping: method and application to endometrial cancer. *BMC Med. Genomics* **4,** 6 (2011).
51. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152,** 633–641 (2013).
52. Webster, J. A. *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.* **84,** 445–458 (2009).
53. Zou, F. *et al.* Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet.* **8,** e1002707 (2012).
54. Colantuoni, C. *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478,** 519–523 (2011).
55. Liu, C. *et al.* Whole-genome association mapping of gene expression in the human prefrontal cortex. *Mol. Psychiatry* **15,** 779–784 (2010).
56. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6,** e1000952 (2010).
57. Zhang, B1. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153,** 707–720 (2013).
58. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
59. Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* **7,** e1002078 (2011).
60. Grundberg, E. *et al.* Population genomics in a disease targeted primary cell model. *Genome Res.* **19,** 1942–1952 (2009).
61. Kabakchiev, B. & Silverberg, M. S. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology* **144,** 1488–1496 (2013).
62. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8,** e1003029 (2012).
63. Ding, J. *et al.* Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *Am. J. Hum. Genet.* **87,** 779–789 (2010).
64. Huang, R. S. *et al.* Population differences in microRNA expression and biological implications. *RNA Biol.* **8,** 692–701 (2011).
65. Rantalainen, M. *et al.* MicroRNA expression in abdominal and gluteal adipose tissue is associated with mRNA expression levels and partly genetically driven. *PLoS ONE* **6,** e27338 (2011).
66. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
67. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** D930–D934 (2012).
68. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 (2012).
69. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
70. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–1048 (2010).
71. Colhoun, H. M. *et al.* Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. *Lancet* **364,** 685–696 (2004).

Acknowledgements

Data from the Heart Protection Study have been contributed by Heart Protection Study Collaborative Group[6].

PROSPER/PHASE: The Prospective Study of Pravastatin in the Elderly at Risk (PROSPER) trial was supported by an investigator-initiated grant from Bristol-Myers Squibb, USA. The study was conducted, analysed and reported independently of the company. The GWAS project PHASE has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement HEALTH-F2-2009-223004. A part of the genotyping was funded by The Netherlands Consortium for Healthy Ageing (NGI: 05060810). Professor Dr J.W.J. is an established clinical investigator of The Netherlands Heart Foundation (2001 D 032).

ASCOT: The Anglo-Scandinavian Cardiac Outcomes Trial (ASCOT) trial was funded by an investigator-initiated grant from Pfizer, USA. The study was investigator led and was conducted, analysed and reported independently of the company. The Genomewide Association Scan was funded by the National Institutes for Health Research (NIHR) as part of the portfolio of translational research of the NIHR Biomedical Research Unit at Barts and the NIHR Biomedical Research Centre at Imperial College, the International Centre for Circulatory Health Charity and the Medical Research Council through G952010. On behalf of the ASCOT investigators, we thank all ASCOT trial participants, physicians, nurses, and practices in the participating countries for their important contribution to the study.

CARDS: We thank the other investigators, the staff and the participants of the CARDS study. A full list of CARDS investigators can be found in original CARDS paper[71]. CARDS was funded by grants to the Universities of London and Manchester by Pfizer, Diabetes UK and the Department of Health.

PARC: This research was supported by the National Institutes of Health: grant U19 HL069757 from the National Heart, Lung, and Blood Institute, and grant UL1TR000124 from the National Center for Advancing Translational Sciences.

TNT: The TNT study was funded by Pfizer, who also provided support for genotyping.

AGES: This study has been funded by NIH contract N01-AG-1-2100, the NIA Intramural Research Program, Hjartavernd (the Icelandic Heart Association) and the Althingi (the Icelandic Parliament). The study is approved by the Icelandic National Bioethics Committee, VSN: 00-063. The researchers are indebted to the participants for their willingness to participate in the study.

ARIC: The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. We thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

BioVU: BioVU receives support through the National Center for Research Resources UL1 RR024975, which is now the National Center for Advancing Translational Sciences, 2 UL1 TR000445. Genotyping was supported via grant U01-HG04603 from the National Human Genome Research Institute and RC2-GM092318 from the National Insitute of General Medical Sciences.

CHS: This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083 and N01HC85086; and NHLBI grants HL080295, HL087652, HL105756 and HL103612 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through AG023629 from the National Institute on Aging (NIA). A full list of CHS investigators and institutions can be found at http://www.chs-nhlbi.org/pi.htm. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

Framingham HS: The Framingham Heart Study work was supported by the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine (Contract No. N01-HC-25195), its contract with Affymetrix, Inc. for genotyping services (Contract No. N02-HL-6-4278) and based on analyses by Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project. A portion of this research was conducted using the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center, also supported by R01HL103612 (PI psaty, subcontract PI, Vasan).

GoDARTS: We are grateful to all the participants who took part in this study, to the general practitioners, to the Scottish School of Primary Care for their help in recruiting the participants and to the whole team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The Wellcome Trust provides support for Wellcome Trust United Kingdom Type 2 Diabetes Case Control Collection (GoDARTS) and informatics support is provided by the Chief Scientist Office. The Wellcome Trust funds the Scottish Health Informatics Programme, provides core support for the Wellcome Trust Centre for Human Genetics in Oxford and funds the Wellcome Trust Case Control Consortium. This research was specifically funded by Diabetes UK (07/0003525) and the Wellcome Trust (084727/Z/08/Z, 085475/Z/08/Z and 085475/B/08/Z). Genotyping of the GoDARTS samples (sample 2) was also funded as part of the EU IMI-SUMMIT programme. K.Z. holds a Henry Wellcome Post-Doctoral Fellowship.

Health ABC: The Health ABC study was supported by NIA contracts N01AG62101, N01AG62103 and N01AG62106. The genome-wide association study was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences and genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging.

HVH: This Heart and Vascular Health Study research was supported by NHLBI grants HL085251, HL073410, HL085251 and HL068986.

MESA: The Multi-Ethnic Study of Atherosclerosis (MESA) and MESA SNP Health Association Resource (SHARe) are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts N01 HC-95159 through N01-HC-95169 and RR-024156. Additional funding was supported in part by the Clinical Translational Science Institute grant UL1RR033176 and is now at the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124. We thank the other investigators in the Pharmarcogenetics Working Group, the staff and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

JUPITER: Genetic analysis in JUPITER was supported by a research grant from AstraZeneca to D.C. and P.M.

Rotterdam study: The Rotterdam Study is supported by the Erasmus Medical Center and Erasmus University Rotterdam; the Netherlands Organization for Health Research and Development (ZonMw); the Research Institute for Diseases in the Elderly; the Ministry of Education, Culture and Science; the Ministry of Health Welfare and Sports; and the European Commission and Municipality of Rotterdam. This work was supported by the Netherlands Genomics Initiative (NGI) Netherlands Organization for Scietific Research (NOW; 050-060-810).

Author contributions

I.P., S.T., H.A.D., M.R.B., X.L., H.R.W., D.I.C., K.Z., B.J.A., B.M.P., G.H., R.M.K., J.W.J. and M.J.C. constituted the writing and analysis group. I.P., S.T., H.A.D. and K.Z. performed quality control on the individual study summary results. I.P. and S.T. performed meta-analysis. I.P., H.A.D., M.R.B., X.L., H.R.W., D.I.C. and R.M.K. performed additional analyses. All analysis and writing group authors extensively discussed the analysis, results, interpretation and presentation of results. All authors contributed to the research and reviewed the manuscript.

Study concept and design of contributing studies by (PROSPER) J.W.J., D.J.S., B.M.B., I.F., N.S. and R.G.J.W.; (ASCOT) M.J.C., P.S., N.P., A.S., D.C.S. and E.O.; (CARDS) H.A.D., H.M.C., P.M.M., J.B., P.N.D., A.D. and G.H.; (PARC) X.L., Y.-D.I.C., J.I.R. and R.M.K.; (TNT) J.J.P.K.; (AGES) L.J.L., T.B.H. and V.G.; (ARIC) C.L.A., E.A.W., T.S., E.B. and C.M.B.; (BioVU) Q.F., W.-Q.W., R.A.W. and J.C.D.; (CHS, HVH) N.S., K.R., T.L., J.I.R., B.M.P. and S.R.H.; (FHS) L.A.C. and V.R.; (GoDARTS) C.N.A.P. and H.M.C.; (HABC) Y.L.; (MESA) X.G., S.R.H., W.P. and J.I.R.; (Rotterdam Study) C.E.d.K., B.H.S., A.G.U., A.H. and F.R.; and (JUPITER) D.I.C., B.J.B., F.N. and P.M.R.

Phenotype data acquisition of contributing studies by (PROSPER) J.W.J., D.J.S., B.M.B., I.F., A.J.M.d.C., N.S. and R.G.J.W.; (ASCOT) M.J.C., P.B.M., P.S., N.P., A.S., D.C.S., E.O. and S.S.H.; (CARDS) H.A.D., H.M.C., P.M.M., J.B., P.N.D., A.D. and G.H.; (PARC) X.L., Y.-D.I.C., J.I.R. and R.M.K.; (TNT) J.J.P.K.; (AGES) G.E.; (ARIC) C.M.B.; (BioVU) W.W.; (CHS, HVH) K.L.W., J.C.B., A.M.A., N.L.S., B.M.P. and S.R.H.; (FHS) L.A.C., C.J.O., V.R.; (GoDARTS) C.N.A.P., L.A.D., K.Z., A.D., A.M. and H.M.C.; (HABC) D.M.H. and S.B.K.; (MESA) W.P. and J.I.R.; (Rotterdam Study) C.E.K., B.H.S., A.H. and O.H.F.; (JUPITER) D.I.C., F.G., J.G.M. and P.M.R.

Genotype data acquisition of contributing studies by (PROSPER) S.T., J.W.J., A.J.M.C. and P.E.S.; (ASCOT) M.J.C., P.B.M., P.S., A.S. and S.S.H.; (CARDS) H.A.D., H.M.C., P.M.M., P.N.D., A.D. and G.H.; (PARC) Y.-D.I.C., J.I.R., D.A.N. and J.D.S.; (TNT) B.J.A., M.P.D., S.M.B., G.K.H. and J.-C.T.; (AGES) A.V.S.; (ARIC) E.B.; (BioVU) Q.F., J.C.D., C.T.L. and F.S.; (CHS, HVH) G.L., J.C.B., K.D.T., J.I.R., K.R., T.L. and S.R.H.; (FHS) C.J.O.; (GoDARTS) C.N.A.P., K.Z., A.D., F.C., H.M.C., M.I.M., L.G., E.A. and WTCCC2; (HABC) Y.L.; (MESA) K.D.T. and J.I.R.; (Rotterdam Study) A.G.U. and F.R.; and (JUPITER) D.I.C., F.G., B.J.B., F.N. and P.M.R.

Primary analysis from contributing studies by (PROSPER) I.P., S.T., A.J.M.C. and P.E.S.; (ASCOT) M.J.C., M.R.B. and H.R.W.; (CARDS) H.A.D., H.M.C. and P.M.M.; (PARC) X.L., Y.-D.I.C. and J.I.R.; (TNT) B.J.A., M.P.D., S.M.B., G.K.H. and J.-C.T.; (AGES) A.V.S.; (ARIC) C.L.A., E.A.W. and T.S.; (BioVU) Q.F., W.W., C.T.L. and F.S.; (CHS, HVH) K.L.W. and G.L.; (FHS) L.A.C., P.G. and J.S.N.; (GoDARTS) C.N.A.P., L.A.D., K.Z. and H.M.C.; (HABC) D.S.E., J.M.S. and J.D.; (MESA) K.D.T., X.G., X.L. and J.I.R.; (Rotterdam Study) C.E.K. and B.H.S.; and (JUPITER) D.I.C., A.Y.C., F.G., J.G.M. and P.M.R.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: B.M.P. serves on the Data and Safety Monitoring Board of a clinical trial funded by the device manufacturer (Zoll LifeCor). N.P. and A.S. received funding from Pfizer for the extended follow-up of the ASCOT UK participants. D.I.C. and P.M.R. received research support for independent genetic analysis in JUPITER from AstraZeneca. F.N. and B.J.B. have employment, stock and stock options in AstraZeneca, a for-profit company engaged in the discovery, development, manufacture and marketing of proprietary therapeutics such as rosuvastatin, but do not consider that this creates any conflict of interest with the subject–matter of this publication. R.M.K. serves on the Merck Global Atherosclerosis Advisory Board. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Postmus, I. *et al.* Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nat. Commun.* 5:5068 doi: 10.1038/ncomms6068 (2014).

Membership of Wellcome Trust Case Control Consortium

Management Committee: Peter Donnelly[77,78], Ines Barroso[79], Jenefer M. Blackwell[80,81], Elvira Bramon[82], Matthew A. Brown[83], Juan P. Casas[84,85], Aiden Corvin[86], Panos Deloukas[79], Audrey Duncanson[87], Janusz Jankowski[88,89,90], Hugh S. Markus[91], Christopher G. Mathew[92], Colin N.A. Palmer[93], Robert Plomin[94], Anna Rautanen[77], Stephen J. Sawcer[95], Richard C. Trembath[92], Ananth C. Viswanathan[96], Nicholas W. Wood[97].

Data and Analysis Group

Chris C.A. Spencer[77], Gavin Band[77], Céline Bellenguez[77], Colin Freeman[77], Garrett Hellenthal[77], Eleni Giannoulatou[77], Matti Pirinen[77], Richard Pearson[77], Amy Strange[77], Zhan Su[77], Damjan Vukcevic[77], Peter Donnelly[77,78].

DNA, Genotyping, Data QC and Informatics Group

Cordelia Langford[79], Sarah E. Hunt[79], Sarah Edkins[79], Rhian Gwilliam[79], Hannah Blackburn[79], Suzannah J. Bumpstead[79], Serge Dronov[79], Matthew Gillman[79], Emma Gray[79], Naomi Hammond[79], Alagurevathi Jayakumar[79], Owen T. McCann[79], Jennifer Liddle[79], Simon C. Potter[79], Radhi Ravindrarajah[79], Michelle Ricketts[79], Matthew Waller[79], Paul Weston[79], Sara Widaa[79], Pamela Whittaker[79], Ines Barroso[79], Panos Deloukas[79].

Publications Committee

Christopher G. Mathew[92], Jenefer M. Blackwell[80,81], Matthew A. Brown[83], Aiden Corvin[86], Mark I. McCarthy[98], Chris C.A. Spencer[77].

[77]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, UK; [78]Department of Statistics, University of Oxford, Oxford, UK; [79]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; [80]Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, 100 Roberts Road, Subiaco, Western Australia, Australia; [81]Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge, UK; [82]Department of Psychosis Studies, NIHR Biomedical Research Centre for Mental Health at the Institute of Psychiatry, King's College London and The South London and Maudsley NHS Foundation Trust, Denmark Hill, London, UK; [83]University of Queensland Diamantina Institute, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, Australia; [84]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK; [85]Department of Epidemiology and Public Health, University College London, London, UK; [86]Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland; [87]Molecular and Physiological Sciences, The Wellcome Trust, London, UK; [88]Centre for Digestive Diseases, Queen Mary University of London, London, UK; [89]Digestive Diseases Centre, Leicester Royal Infirmary, Leicester, UK; [90]Department of Clinical Pharmacology, Old Road Campus, University of Oxford, Oxford, UK; [91]Clinical Neurosciences, St George's University of London, London, UK; [92]King's College London, Department of Medical and Molecular Genetics, School of Medicine, Guy's Hospital, London, UK; [93]Medical Research Institute, University of Dundee, Ninewells Hospital and Medical School, Dundee, UK; [94]King's College London Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Denmark Hill, London, UK; [95]University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge, UK; [96]NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK; [97]Department of Molecular Neuroscience, Institute of Neurology, Queen Square, London, UK; [98]Oxford Centre for Diabetes, Endocrinology and Metabolism (ICDEM), Churchill Hospital, Oxford, UK.